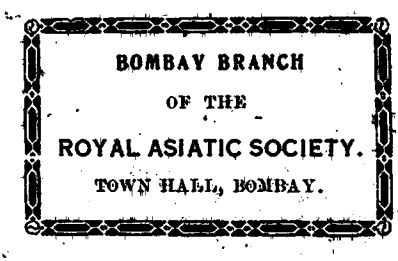
Damaged books
physically verified
by B.V.P. & M.C.G.
27/6/2000
Lot No:-38

310
BOW/Elc
292246



Digitized with financial assistance from the
Government of Maharashtra
on 26 June, 2018

STUDIES IN ECONOMICS AND POLITICAL SC

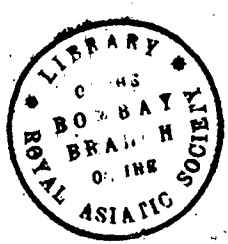
Edited by Prof. W. A. S. HEWINS, M.A.,

Director, London School of Economics.

ELEMENTS

OF

STATISTICS



PART I.

CHAPTER I.

SCOPE AND MEANING OF STATISTICS.

00029224

ELEMENTS

OF

STATISTICS

BY

ARTHUR L. BOWLEY, M.A., F.S.S.,

Lecturer in Statistics at the London School of Economics and Political Science;

Author of "Wages in the United Kingdom in the Nineteenth Century."

SECOND EDITION.

29224

LIBRARY OF THE ROYAL STATISTICAL SOCIETY

London:
P. S. KING & SON,
ORCHARD HOUSE, WESTMINSTER.
New York: CHARLES SCRIBNER'S SONS.
1902.

PREFACE.

THIS book is based on lectures given at the London School of Economics and Political Science in the five years following its foundation in 1895. There seems to be no text-book in English dealing directly and completely with the common methods of statistics. English writings on the various branches of the science are for the most part in the form of articles in the journals of learned societies. Professor Mayo Smith in his *Statistics and Sociology* proceeds almost at once to historical applications; while in Professor Meitzen's *Geschichte, Theorie, und Technik der Statistik*, issued in English by the American Academy of Political and Social Science, so much space is devoted to the history of the development of statistics, and the book is so slight, in comparison with the wide field it covers, that many elementary methods are treated very cursorily. In the excellent books in French, German, and Italian on this subject there is a general tendency to deal at length with the history of official statistics, the limits of the science, and particular applications of the theory of probability, to the exclusion of more general matter; so that a student must refer to the works of Dr Mayr, Professor Westergaard, Professor Lexis, Professor Gabaglio, M. Block, and Dr Bertillon before he is completely ac-

quainted with the elementary methods of statistics. The result is that there is no compact statement of principles acknowledged by statisticians, of the methods common to most branches of statistical work, of the artifices developed for handling and simplifying the raw material, and of the mathematical theorems by the use of which the results of investigations may be interpreted. This book forms an attempt to supply this want, so far as can be done without undue length. No place has been given in it to the history of statistics, and it does not contain any summary of the main groups of statistics extant; several tables, drawn from a wide range of subjects, are given, but only to illustrate particular methods, and their choice has been determined by their suitability for this purpose. In the chapter on Collection of Material some account is offered of the genesis of the most important English statistics: the great part of the figures tabulated in the *Statistical Abstract* can be traced back to the householder's schedule of the Population Census or the custom house returns of foreign trade, while the chief statistics accessible for the study of modern social questions have come from the Wage Census of 1886 or are collected by the Labour Department: it is hoped that the account of these four groups of figures will afford some help in judging of their accuracy and limitations. Considerable space has been allotted to the subjects of Averages and Diagrams, because their use is universal, and, while their principles and technique are simple, their application is often misunderstood. The chapter on Accuracy is based on the Newmarch Lectures of 1897, and may perhaps be found to contain something that is new, a claim which is not made for the great part of the book. The treatment

throughout is intended to be suitable for those whose mathematics have not been carried to any height or have become rusty from disuse. With this view, when mathematical symbols were unavoidable, the preliminary hypotheses have been first discussed without algebraic notation and at some length, and those proofs have been chosen which require the minimum mathematical knowledge rather than those which lead most directly to the result. Thus the most important results of the Theory of Error have been obtained without the use of the Differential or Integral Calculus, and it is hoped that the greater part even of the chapter on Correlation will be intelligible to those who are not so well equipped as the Major-General in the *Pirates of Penzance*. Part II. is intended to be introductory and is certainly incomplete; the normal law of frequency is the only one discussed, and the correlation of three variables is untouched. The more advanced treatment of this part of the subject is likely to be of interest to but few, who will have little trouble in obtaining the books and journals in which the further development may be found. Short bibliographies are added to the chapter on Interpolation and to Part II. for this purpose. It is hoped that this elementary handling may be of use to some who are interested in the statistical arguments based on the Laws of Probability, and that the definitions, formulæ, and proofs given may save others from the necessity of searching in books, long out of print, for elementary theorems and deductions. The treatment in Part II., Section II., is peculiar in that it leaves very much in the background the Method of Least Squares; the phrase, useful in some connections, seems to make the application of the Law of Error to statistics

unnecessarily complex. I am much indebted to Professor Edgeworth, who has not only given me continued help both privately and by his publications in the study of the mathematical treatment of statistics, but has also read Part II. in proof and suggested many useful and important alterations. My thanks are also due to Professor Everett and Mr W. F. Sheppard for help in the chapter on Interpolation, and to Mr C. P. Sanger and Mr H. Clissold for reading great parts of the book in proof.

A. L. B.

London School of Economics,
January 1901.

NOTE TO SECOND EDITION.

It has not been possible, even if it were advisable, to introduce any considerable changes in this edition. The text has been revised and notes have been added where the meaning had proved obscure. Pages 117, 118, relating to the importance of weights, have been re-written, and the canon that errors in weighting are generally unimportant has been explained and qualified. The measurement of dispersion has been developed on pages 126 and 136, and an Appendix has been added dealing with the asymmetrical curve of error on the lines laid down by Professor Edgeworth. It is hoped that the views of friendly critics have by these means been met, and that the book has been rendered more complete.

A. L. B.

September 1902.

TABLE OF CONTENTS.

PART I.

CHAP.		PAGE
I.	SCOPE AND MEANING OF STATISTICS	3
II.	THE GENERAL METHOD OF STATISTICAL INVESTIGATION	17
III.	ILLUSTRATIONS OF METHOD—	
	SECTION I. The Population Census	23
	„ II. The Wage Census	33
	„ III. The Work of the Labour Department	41
	„ IV. The Statistics of England's Foreign Trade	63
IV.	TABULATION: General—Mr Booth's Use of Census—Agricultural Earnings—U.S.A. Wage Statistics—Wage Census—Changes of Wages	73
V.	AVERAGES: *A.* Simple; *B.* Weighted; *C.* The Mode; *D.* The Median; *E.* The Geometric Mean; *F.* Statistical Coefficients; *G.* General	107
VI.	USE OF AVERAGES IN TABULATION	133
VII.	THE GRAPHIC METHOD—	
	SECTION I. General Purpose	143
	„ II. Historic Diagrams	159
	„ III. Comparisons of Series of Figures	168
	„ IV. Periodic Figures	178
	„ V. Logarithmic Curves	188
VIII.	ACCURACY	199
IX.	INDEX-NUMBERS	217
X.	INTERPOLATION—	
	SECTION I. General	233
	„ II. Algebraic Treatment	242

PART II.

APPLICATION OF THE THEORY OF PROBABILITY TO STATISTICS—		
SECTION	I. Introductory	261
„	II. The Equation of the Curve of Error	269
„	III. To what Groups does the Law of Error apply?	293
„	IV. The Permanence of Certain Small Numbers	301
„	V. Extension of the Law of Error and Applications	303
„	VI. The Theory of Correlation	316
APPENDIX		329
INDEX		335

LIST OF DIAGRAMS.

	FACING PAGE
Graphic Method of Finding the Median, Quartiles, and Deciles	128
Graphic Representation of Wage Statistics	145
Total Value of British and Irish Produce Exported from the United Kingdom, 1855-1898	151
Graphic Method of Determining the Median and Modes	155
Revenue of the United Kingdom, 1850-1898	159
Importation of Wheat and Wheat Flour, 1862-1898	164
History of the Cotton Trade, 1854-1898	165
Trade with British Possessions and Foreign Countries	172
Marriage Rate and Foreign Trade	175
Fluctuations of Employment, 1855-1893	179
Growth of Imports and Exports (Natural and Logarithmic Scale)	190
Marriage Rate and Employment (*Logarithmic* Scale)	193
Symmetry of the Binomial Curve near its Greatest Ordinate	277
Table of Minima and Maxima Temperatures	323
Example of the Skew Curve of Error	334
The Curve of Error	*At end*

CHAPTER I.

SCOPE AND MEANING OF STATISTICS.

VERY many definitions have been given of the word *statistics*, and each author who has written on the subject has assigned new limits to the field which should be included in its scope. It will not be necessary for the purpose of this book to discuss the merely verbal differences involved, but only to explain what is intended by its title, and to consider the limits of the science which it is proposed to investigate. It will be useful, however, to mention some possible definitions.

Definitions of statistics.

Statistics may, for instance, be called *the science of counting*. Counting appears at first sight to be a very simple operation, which any one can perform or which can be done automatically; but, as a matter of fact, when we come to large numbers, *e.g.*, the population of the United Kingdom, counting is by no means easy, or within the power of an individual; limits of time and place alone prevent it being so carried out, and in no way can absolute accuracy be obtained when the numbers surpass certain limits. Great numbers are not counted correctly to a unit, they are estimated; and we might perhaps point to this as a division between arithmetic and statistics, that whereas arithmetic attains exactness, statistics deals with estimates, sometimes very accurate, and often sufficiently so for their purpose, but never mathematically exact. Statistics generally relate to numbers so great that their estimation is beyond the power of an individual, and requires the co-operation of an organised body of workers. Though the collection of numbers by several persons and the mere addition of the results seem simply questions of arithmetic, yet in practice two difficulties soon occur. First, it is not easy to define the thing to be counted so explicitly that all the tellers shall admit and reject instances on the same principles; for such

The science of counting.

Distinction between statistics and arithmetic.

Statistics as co-operative counting.

simple objects as the number of rooms or stories of a house, a person's age, even an individual, give rise to such complex questions of definition that it is often impossible to tell from a short description of a category exactly what items are included in it. Secondly, numerical errors cannot be avoided when many workers are involved; for some among a large number of persons will be inaccurate, some unintelligent, some will not obtain complete information, and when their reports are compiled there will be occasional mistakes in copying and errors in tabulation. A total which is the result of the work of many hands will certainly from one cause or another fall short of complete accuracy. But though all estimates of this nature are sometimes included under the term *statistics*, this definition at once is too wide, and also does not bring out the distinctive nature of statistical method.

It is better, in fact, to define statistics *a posteriori*. In dealing with masses of figures, large numbers descriptive of groups, series of totals or averages relating to different dates or places, it is found that special methods become necessary—methods which depend on particular properties of large numbers, methods which are suitable for describing complex groups so that they can be easily comprehended, methods for analysing the accuracy of statements, for measuring the significance of differences, for comparing one estimate with another. Those estimates to which these methods apply are within the scope of statistics; it is the study of these methods that is the object of this book. It is clear that, under our tentative definition, statistics is not merely a branch of political economy, nor is it confined to any one science. A knowledge of statistics is like a knowledge of foreign languages or of algebra: it may prove of use at any time under any circumstances.

Statistics as a method.

Generality of statistical method.

It may be interesting to trace the connection of statistical method with various branches of knowledge. To begin with the physical sciences: there are two points in which this method touches astronomy. The method of least squares was introduced by an astronomer, anxious to choose the best of several slightly discrepant observations of the position of a star. In most physical observations several measurements are taken of the same quantity, and it is found that, however carefully they are made, they never absolutely

Its use in the physical sciences.

agree; just as the averages obtained by different statisticians from the same series of sociological observations are generally not identical. From such a group of measurements it is necessary to deduce the most probable estimates; this is done by the application of the law of error, known as the method of least squares.

The other point of resemblance of statistical to astronomical method is common also to geology and to most applied sciences.

Progressive accuracy.

The course of scientific measurement has generally been to take first a rough observation of a quantity, such as the distance of the sun, the thickness of a stratum, the atomic weight of an element, the specific gravity of a substance; then, as information accumulated, as the precision of instruments increased and methods were better adapted, to make the measurement gradually more and more accurate. It is important to appreciate this development, for in the present state of our knowledge, many statistical measurements cannot be made with precision for want of data, and a critic is inclined to say that for this reason preliminary estimates are valueless; but from the scientific point of view this criticism is wrong, for a faulty measurement made on logical principles is better than none, and may lead to others with progressive improvement.

Passing by the general resemblance of statistical investigations to all scientific experiments, we may notice the use of statistics

Statistics and biology.

in biology. It was, perhaps, not recognised before the publication of Professor Karl Pearson's investigations,* that the whole doctrine of evolution and heredity rests in reality on a statistical basis. It is in this direction that the most important new work in statistics is being done. It may be worth while to sketch very briefly the nature of the problem. Out of a great number of observations, say the measurements of the heights of a group of men, the type is found—the average, about which all the measurements are grouped according to some definite law. The problem is then to determine whether this type or the grouping about it changes, and in what way. The differences found in successive generations form the data on which arguments as to evolution and development are founded. The method applies equally to fossil remains, to zoological species, and to many other groups. If it is neglected, many

* See *The Grammar of Science*, chap. x. *seq.*, and the references there given.

valid arguments lose a great part of their force, and theories are founded on personal impressions of phenomena instead of on scientific measurement. The work done in this direction becomes of immediate use to the student of social questions. The average wage and the grouping about it and the change in these quantities present precisely similar problems; the change in the purchasing power of money is calculated by the same mathematical formulæ; in fact, these methods furnish the only accurate way of measuring numerical changes in complex groups. Much valuable information has been collected in anthropometrical laboratories, which has increased the statistician's knowledge of facts and given birth to important theoretical principles.

Statistics and meteorology.

Meteorology has much in common with statistics. The chief measurements taken for the purposes of this science are of temperature, barometrical pressure, moisture of the air, and force of the wind. One of the problems attacked is again that of finding the type from a group of observations, and of measuring its change. The tables which state the average temperature year by year are in many ways similar to those which the Registrar-General publishes of births, deaths, and marriages. Without the aid of statistical method, the averages obtained show mere numbers from which no logical deductions can be made. With the help of this knowledge, it can be seen whether the change from year to year is significant or accidental; whether the figures show a progressive or periodic change; whether they obey any law or not. The problem is easily seen to be of importance for forecasting the future population and for many similar purposes.

We are thus brought by a short step to the province to which statistics has sometimes been confined: the study of demography.

Statistics and demography.

If in demography we include, not merely the measurement of the numbers of the population, the birth, marriage, and death rates, the distribution by age, by sex, and by locality, in fact, the figures which naturally come from the census and the Registrar-General's returns; but include also, industrial and social measurements, of distribution of the population by trade, of income, wages, production, foreign trade transportation, and so forth; we have extended the limits of demography till it includes the majority of the statistical investigations directly interesting to students of sociology

or of political economy. Without stopping to decide the exact limits of demography, we can quickly pass to another definition of statistics (so far as it concerns such students) on which it is wished to lay a certain stress: *statistics is the science of the measurement of the social organism, regarded as a whole, in all its manifestations.* In a monograph, after the fashion of Le Play, a single family is studied; the occupations and earnings of its members, the way these earnings are spent, and its economic position generally are set down; but this study is not so far statistical. In demography we study the same quantities when groups of families are concerned; the number of families engaged in certain industries, and their average receipts, expenditure, and savings; here we have statistics. In the monographic method the individual is everything; in the statistical method, nothing. When we wish to obtain a measurement of the group, peculiarities of individuals receive no attention; it is only when the same peculiarities are possessed by many persons that they become of importance. Statistics may rightly be called the science of averages. In the measurement of a complex group, say of incomes and wages, the exceptional artiste who can earn £100 in an evening, and the inefficient labourer who can only make sixpence a day, affect only slightly the general average; they are not entered in separate categories; but the large group of skilled artisans who can earn over forty shillings a week, or of casual labourers who make less than fifteen shillings, are entitled to separate notice. The exact specification to be adopted is only a question of degree, which differs with the nature of the particular investigation in hand. The object of a statistical estimate of a complex group is to present an outline, to enable the mind to comprehend with a single effort the significance of the whole. To do this it is necessary to exclude rigorously any presentation of details, for the same reason that, in a painter's rendering of a tree, the individual leaves are not distinguished. The outline will be a little blurred, a little inaccurate; but it will be as distinct and detailed as the mind has power to grasp it, or the eye to see it; the impression will be rightly given. There is a very important principle involved in this method. The individual members of a group vary continually, the whole group varies very slowly. It is impossible to follow or measure the motions of separate atoms; it is comparatively easy to state the laws of

Statistics relate to the social organism as a whole.

motion for a solid body. Great numbers and the averages resulting from them, such as we always obtain in measuring social phenomena, have great inertia. The total population, the total income, the birth and death rates, average wages, change very little; similar quantities relating to a single family change very fast. It is this constancy of great numbers that makes statistical measurement possible. It is to great numbers that statistical measurement chiefly applies.

Statistics and political economy.

The relation of statistics to political economy is a simple one. Professor Marshall says,* "Statistics are the straw out of which I, like every other economist, have to make the bricks." The statistician furnishes the political economist with the facts, by which he tests his theories or on which he bases them. Since the economist deals chiefly with phenomena relating to groups, and regards the individual only as a member of a group, it is to statistics as the science of averages that he looks for his information. When he is dealing with national economy, with the volume of trade, for instance, or the purchasing power of money, he is limited to pure theory, till statistics as the science of great numbers has provided the facts. The chemist experimenting in his laboratory is like the statistician; the chemist theorising in his study is like the economist. Because of this relation it may be held to be the business of the statistician to collect, arrange, and describe, like a careful experimentist, but to draw no deductions; even in an investigation relating to cause and effect, to present evidence but not conclusions. As a distinct operation, of course, the statistician may assume the rôle of the economist, for the same man may well be fitted to conduct the experiment and fit the theory. And just as a theoretical chemist will have little or no power unless he fully appreciates experimental methods and difficulties, even if he has not the manual dexterity to conduct them to perfection himself, so no student of political economy can pretend to complete equipment unless he is master of the methods of statistics, knows its difficulties, can see where accurate figures are possible, can criticise the statistical evidence, and has an almost instinctive perception of the reliance that he may place on the estimates given him.

The proper function, indeed, of statistics is to enlarge indi-

* Evidence to the Committee on the Census, 1890.

vidual experience. An individual is limited to what he can himself see, a very small part of one division of the social organism; his knowledge is extended in various ways, by the conversation of his acquaintance, by newspaper reports, by the writings of experts. According to his ability and power of judgment, he will be able to form a correct view of the numerical importance of groups of persons and things; but it is in the highest degree improbable that he will not have been biassed by the peculiarities of his position, and that he will place his different items of information in the right perspective; and he will not be able to gauge rightly the accuracy of his data. As soon as he begins to examine these points he is undertaking a statistical investigation, and will very soon find himself involved in all the difficulties and problems from which a knowledge of statistical method alone can disentangle him. This is the obvious answer to those who deny the use of statistics. A statistical estimate may be good or bad, accurate or the reverse; but in almost all cases it is likely to be more accurate than a casual observer's impression, and in the nature of things can only be disproved by statistical methods.

Statistics *versus* individual experience.

A chief practical use of statistics is to show relative importance, the very thing which an individual is likely to misjudge. Statistics are almost always comparative. The absolute magnitude of a quantity is of little meaning to us till we have some similar quantity with which to compare it. A statement of the number of paupers in the United Kingdom is valueless unless we know the total population. A statement of the number of gallons of water supplied per head to the people of East London is of little meaning to us till we know the quantity supplied to other towns. The average wage, shown in the Wage Census, does not convey its full significance till we have similar computations for other countries or relating to other years. In the case of most statistical estimates, it will be found that we need another for comparison before we can appreciate the meaning of the first.

Statistics are comparative.

If the group of objects which we wish to measure is large, its enumeration will be beyond our unassisted efforts, or those of any organisation at our command. Some investigations, indeed, have been successfully conducted by private organisations, for instance, those which resulted in Mr Booth's *Life and Labour of the People*, and Leone Levi's *Wages and*

Official statistics:

Earnings; but in general the measurement of a part of the social body or industrial organism must be undertaken by the central or local governments, if it is to be successfully carried out. The fact that this is the case explains the heterogeneousness and imperfection of the mass of statistics extant. A government naturally collects numerical information only in relation to its own functions. Thus the administration must know the numbers of the population and the area of the country in gross and in detail for its own purposes. Large groups of figures come simply from the necessity of public account-keeping. Many official figures are bye-products; for office purposes an account is kept of all transactions in which the government has a hand, and of industries subject to special regulations; and the government publishes most of the figures which thus come in its way. To such causes are due our knowledge of the statistics of income, education, imports, railways, mines, factories, and so on. Some such publications are only survivals from a former time, when the figures were directly needed, such as Gazette wheat prices (still used for the calculation of tithes), and to some extent, statistics of exports. Though few figures are collected simply for scientific purposes, yet in many cases schedules issued for administrative ends are used at the same time for the reception of other information, of use chiefly to the sociological student; much of the Census information comes under this heading. A view of those figures, relating to the United Kingdom, which are easily accessible to the student, can be obtained by turning through the annual *Statistical Abstract for the United Kingdom*, the *Annual Abstract of Labour Statistics*, and the *Registrar-General's Annual Report;* in one or other of these, summaries of, and references to, most official statistics are to be found.

their incompleteness.

It is clear that figures collected simply in connection with administrative purposes are not likely to be precisely those which are needed by the student of sociology or political economy. Even where the wants of the official and the student are nearly identical, the classification and tabulation may not meet scientific requirements. There has, indeed, been considerable progress in recent years, due one may suppose to the influence of Sir R. Giffen, in the direction of amassing statistical information not absolutely needed by the administration, and most of the work of the Labour Department is of this kind; but very much more might reasonably be done,

at an expense which would be almost negligible when considered in relation to the national income. Thus the census might be made, in part at least, quinquennial, and the body of workers, who are organized once in ten years to conduct it, only to be disbanded when the report is issued, might be made permanent and entrusted with the organization of a decennial industrial census. Market prices of many staple commodities could be tabulated by local officials in the same way as wheat sales are now registered. Movements of goods by rail could be tabulated in the same way as transport by water, and the anomaly that we know more of our foreign than of our home trade be removed. The production of factories might be returned as well as that of mines. A permanent government office might well be charged from time to time with special investigations, similar perhaps to the Wage Census of the Board of Trade. It needs very little study of statistics or of political economy, to feel the pressing need of some of this information.

Statistics specially needed.

Attention may be drawn to some of the gaps in our knowledge. When dealing with our national income we can obtain statistics of wages, and of income subject to tax; but for salaries below the exemption limit, and for part of the income received for foreign investments, we are forced to rely on educated guesses. For the change of the purchasing power of money we know, thanks chiefly to the *Economist* and trade newspapers, the course of wholesale prices, but many interesting calculations are brought to a standstill because of the complete dearth of records of retail prices. With regard to wages, we can estimate fairly accurately standard and average wages, but, in default of an industrial census, do not know how many persons are in receipt of each given wage, nor the relative numbers of masters and men. We know fairly well the mass of trade that leaves or reaches our shores, but as regards the far greater mass of our internal trade our ignorance is almost complete. Till there is a public demand for such information, it will need a very enlightened government to spare the time, trouble, and expense necessary for a systematic attempt to fill up these gaps; but we can all do something towards this enlightenment, and in furtherance of this demand, by studying what has been done in other countries, and building up a knowledge of the science of statistical investigation.

The absence of such a demand is perhaps due to a widely

spread and not unreasonable distrust of statistical estimates, crystallized in the common remark that "anything can be proved by statistics." This is to a great extent the fault of the criticising public themselves: they are always requiring and the newspapers always supplying information, which depends on a statistical basis, but for which good statistics are not to be found for one or other of the reasons already indicated. The informant must perforce turn to inaccurate estimates, and the public has no knowledge or discrimination as to what estimates rest on satisfactory data, or indeed as to what quantities are capable of statistical evaluation. Again, figures which cover only part of the subject, such as the Wage Census average, or the *Labour Gazette* returns of unemployed, may be quoted as universal; mere estimates, made for quite other purposes, may be given as accurate and complete; and on such unreliable premises arguments are based, which naturally, by a judicious choice of material, can be made to support any theory at pleasure. It will generally be found that the statistician, on whose authority such statements are supposed to be based, is not to blame. Some of the common ways of producing a false statistical argument are to quote figures without their context, omitting the cautions as to their incompleteness, or to apply them to a group of phenomena quite different to that to which they in reality relate; to take estimates referring to only part of a group as complete; to enumerate the events favourable to an argument, omitting the other side; and to argue hastily from effect to cause, this last error being the one most often fathered on to statistics. For all these elementary mistakes in logic, statistics is held responsible.

Distrust of statistics: its causes.

Perhaps statisticians themselves have not always fully recognised the limitations of their work. At best they can measure only the numerical aspect of a phenomenon; while very often they must be content with measuring not the facts they wish, but some allied quantity. We wish to know, for instance, the extent of poverty, its increase or diminution: poverty we cannot define or measure, and we cannot even count the number of the poor; all we can do is to state the number of officially recognised paupers, and add perhaps some estimates from private sources; but this gives us no clue to the intensity of poverty in individual cases. Or we wish to obtain

Limitations of statistics.

statistics of health: all we can measure is the death-rate and average length of life, and the prevalence of some diseases, very different matters. The statistician's contribution to a sociological problem is only one of objective measurement, and this is frequently among the less important of the data; it is as necessary, however, to its solution as accurate measurements are for the construction of a building.

CHAPTER II.

THE GENERAL METHOD OF STATISTICAL INVESTIGATION.

CHAPTER II.

THE GENERAL METHOD OF STATISTICAL INVESTIGATION.

AT first sight it will seem as if there were no method common to all statistical investigations, and indeed the processes differ so widely that it is not easy to outline a scheme which will include them all; but the following sequence is generally indicated* as of general application, and will serve at least to thread an examination of methods together: (1) the Collection of Material, (2) its Tabulation, (3) the Summary, and (4) a Critical Examination of its results. These processes will be discussed in detail in the following chapters.

Preliminary knowledge necessary or expedient.

It may be well to state what equipment is necessary for the student who wishes to learn statistical methods. In collection and tabulation common-sense is the chief requisite, and experience the chief teacher; no more than a knowledge of the simplest arithmetic is necessary for the actual processes; but since, as we shall see immediately, all the parts of an investigation are interdependent, it is expedient to understand the whole before attempting to carry out a part. For summarising, it is well to have acquaintance with the various algebraic averages, and with enough geometry for the interpretation of simple curves, though all the operations can be performed without the use of algebraic symbols. For criticism of estimates and interpretation of results, it is necessary to use the formulæ of more advanced mathematics, and it is obviously expedient to understand the methods by which these formulæ are obtained to ensure their intelligent use. They are specially necessary for the comparison of complex groups, and for estimating the significance of a divergence from the average, or the deviations in a list of periodic figures.

---

* See, *e.g.*, Dr Bertillon's *Cours élémentaire de Statistique*, to which the present author is indebted for some of the treatment in the following pages.

(1.) Information is generally collected by issuing blank circulars, forms of inquiry, to be filled in either by a few officials or by many individuals, and the proper drawing up of this form is one of the chief tasks in a good investigation. Before this form is issued it is necessary to formulate a complete scheme of the whole undertaking, and even to have some idea of what the resulting figures will be, so as to be able to arrange the details of the organization on the right scale, and adjust the tools used to their purpose. As already pointed out, the object whose measurement is wanted is not in general exactly that which can be measured, and the measurable quantity nearest to it must be found; *e.g.*, when the average annual earnings of the working class were in question, the quantity first measured was the average weekly wage. Then some technical knowledge of the particular subject is needed; and, if not possessed, a preliminary inquiry on a small scale may be necessary to show how to fit means to ends. The people who possess the information required must be discovered and interrogated at first hand. The questions put must be those which will yield answers in a form ready for tabulation, and the scheme of tabulation must therefore be thought out beforehand. The questions must be so clear that a misunderstanding is impossible, and so framed that the answers will be perfectly definite, a simple number, or "yes" or "no." They must be such as cannot give offence, or appear inquisitorial, or lead to partisan answers, or suppression of part of the facts. The mean must be found between asking more than will be readily answered and less than is wanted for the purpose in hand. The form must contain necessary instructions, making mistakes difficult, but must not be too complex. The exact degree of accuracy required, whether the answers are to be correct to shillings or pence, to months or days, must be decided. Every word and every square inch of space must be keenly criticised. A little trouble spent upon the form will save much inconvenience afterwards.

Collection: blank forms; nature of the questions.

(2.) In considering what method is to be adopted for tabulation, we must remember that the investigation is intended to furnish the answers to certain definite questions—how many people, what wage, what price—and each column must present some total which is relevant to these

Tabulation.

questions. The exact scheme employed will differ in different inquiries. In the population census, the tabulation is almost automatic; in the wage census, the best and simplest way to show the grouping about the average wage in each occupation had to be specially devised; in trade statistics the number of different categories to be adopted and the limits of each raise difficult questions. In general, the scheme of investigation requires knowledge of certain groups; and the totals resulting from tabulation should show the numbers of items in these, so that after tabulation, instead of the chaotic mass of infinitely varying items, we have a definite general outline of the whole group in question.

(3.) When the raw material is worked up to this point, skill of a different kind is wanted. From the numbers obtained, we have to pick out the significant figures; so to present the totals and averages as to give a true impression to an inquirer; to summarise briefly the information obtained; to concentrate the mass into a few significant averages, and to describe their exact meaning in the fewest and clearest words, for it is the result of this concentration which will generally be used and quoted. To do this skilfully requires an acquaintance with the method of averages and the use of diagrams. It may further be necessary to fill in unavoidable gaps in the figures in order to supply estimates for intermediate years; this needs a study of the dangerous method of interpolation. Finally, the verbal description of the process, its genesis and results, and an estimate of its accuracy must be added, and then the investigation is complete.

Averaging and summarisation.

(4.) The student who has to make use of statistics should not be content to take the results of an inquiry on authority, but ought to acquaint himself with all these details of method. Before the results can be criticised, it is necessary to know the complete genesis of the figures; whether the whole field was covered; exactly whence the information tabulated was obtained; whether there was a possibility of bias; how nearly the individual answers were correct; whether the informants really knew the facts they related, and if they were likely to state them correctly. The published statement of the results should show clearly the whole scheme of collection so as to make this criticism possible; in particular, specimens of the original blank forms should be

Criticism of results.

included, so that the reader can judge whether the original answers correspond exactly to the form of tabulation employed. Internal evidence often leads to much useful criticism. It can be seen whether the number of returns for each group is proportional to its importance, or if a specially important figure depends on only slight evidence. The continuity of the figures can be examined, and the causes of sudden gaps investigated. The returns can be divided into sample groups, and the extent of the correspondence of these groups to the general result will often indicate whether the returns are sufficiently general. A careful study of the more minute tabulations may show within what percentage the final numbers may be expected to be correct.

The most important function of statistics is to produce evidence showing the relation of one group of phenomena to another; for the information obtained is presumably intended as a guide for action, the guidance is generally needed to show what actions are likely to produce certain desired effects, and this is best investigated by finding how such effects have been produced in the past. We have then to determine whether changes in one measurable quantity (*e.g.*, the duties on corn) have produced changes in another (*e.g.*, the amount of pauperism); a problem generally insoluble, but one on which most light can be obtained by the study of the relevant statistics in the light of mathematics, the mathematics of probability, and it is in this particular branch of mathematics that recent statistical progress has been chiefly made.

Such questions, however important, are somewhat abstruse, and presuppose a certain amount of technical knowledge which is not in the possession of the general student. The plan of this book is to postpone all questions requiring such technical or mathematical knowledge to the Second Part, and to confine our earlier discussions to problems needing no special training or equipment.

CHAPTER III.

ILLUSTRATIONS OF METHOD.

List of the MEMBERS of this FAMILY, of VISITORS, of BOARDERS, and of SERVANTS, who SLEPT or ABODE in this dwelling on the NIGHT of SUNDAY, APRIL 5th 1891.

Cols.	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
	NAME and SURNAME.	RELATION to Head of Family.	CONDITION as to Marriage.	SEX.	AGE last Birthday.	PROFESSION or OCCUPATION.				WHERE BORN.	If (1) Deaf and Dumb. (2) Blind. (3) Lunatic Imbecile or Idiot.
	No persons absent *on the night of Sunday, April 5th, to be entered here;* except *those who may be* travelling *or out at* Work *during that night (and are not elsewhere returned), and who* Return Home on Monday, April 6th. Write after the Name of the Head of the Family, the Names of his Wife, Children, and other Relatives; then Visitors, Boarders, and Servants.	State whether Head, or Wife, Son, Daughter, or other Relative, Visitor, Boarder, or Servant.	Write either "*Married*," "*Widower*," "*Widow*," or "*Single*," opposite the Names of all Persons, except Young Children.	Write "M" opposite Males and "F" opposite Females.	For *Infants under One Year* state the age in *Months*, writing "*Under 1 Month*," "*1 Month*," "*2 Months*" &c.	*Before filling up columns 6, 7, 8, and 9, you are requested to read carefully the Special Instructions printed on the other side.*	Employer.	Employed.	Neither Employer, nor Employed, but working on own account.	Opposite the Names of those born in ENGLAND and WALES, write the COUNTY, and TOWN or PARISH. If born in SCOTLAND, IRELAND, the BRITISH COLONIES or the EAST INDIES, state the *Country* or *Colony*. If BORN IN FOREIGN PARTS, write the particular *State* or *Country*; and if also a BRITISH SUBJECT, add "*British Subject*," or "*Naturalised British Subject*," as the case may be.	Write the precise Infirmity, if any, opposite the name of the person; and if the Infirmity dates from childhood, add "*from childhood.*" Do not use such a general term as "*Afflicted*" or "*Infirm.*"
	If you occupy less than five rooms, write in this space the number of rooms occupied by you.		I declare the foregoing to be a true Return, according to the best of my knowledge and belief. Witness my hand, (*Signature*).....								

To face page 23.

CHAPTER III.

ILLUSTRATIONS OF METHOD.

SECTION I.—THE POPULATION CENSUS.

THE population census will provide good illustrations of the principles laid down in the last chapter, both because we shall be at first on familiar ground, since every one knows its scheme, purpose, and details, and because the form of inquiry used for the collection of the original data brings out very prominently the difficulties met with in detailed statistical investigations.

The census.

The first thing to be considered is the exact object for which the census is undertaken. It is for demographical purposes; to supply information as to the numbers and local distribution of the population, the numbers of each sex and age, their so-called civil condition (*i.e.*, whether single, married, or widowed), and their nationality. This is the minimum information necessary for administrative purposes. In addition to these facts there are very many others which the statesman and the economist wish to know about each member of the population, and the census form is the only means in England of collecting universal data; the question as to which of these shall be investigated and which neglected, is decided more by expediency than on principle. Of these desiderata the following may be mentioned: the size and structure of the family, its position in the social scale, the economic position of its head; the nature of employment of its members, the wage or income of each member and of the family as a whole, the rent and size of their house, their educational condition, the ages at which they commenced or retired from work, their migrations, their combination in religious or other bodies, and their infirmities. It is clear that some of this information must be dispensed with, if the form is not to be overcrowded, and if the tabulation is to be finished in any

Its object.

The choice of questions.

reasonable time; and an examination of the general nature of the questions which can suitably be put will show how the necessary selection is made.

First, the questions must be those which the informant is able to answer. Now, if the questions were only to be put to educated and methodical persons, doubtless a full account could be given of the family migrations and of the ages at which each member had been at work; but the peculiarity of the census is that it is universal, and the questions must be such that the least educated and most unthrifty householder shall be able to answer; in many cases such facts would have been unrecorded and forgotten.

Ability to answer.

Secondly, the questions must be perfectly definite, so that there can be no doubt as to what the right answer should be. The only answers which are of value to the statistician are "yes," "no," or a simple number. Adjectives and adverbs such as many, often, partly, &c., bear different numerical meanings to different people, and, though they may express fairly clearly the position of an individual, are nearly useless for tabulation,* which is their only purpose so far as the census is concerned. Thus the question as to education would have to be, not "state whether well, moderately, or badly educated," but "state at what age school was left," or "how many years at school?" But even if such questions were not excluded by our first test, by the forgetfulness of the informant, the statements given would be of little practical value, and very often incorrect. An inquiry as to wage and income could not be made sufficiently definite without so many questions as to require a form to itself; for wages, as we shall see when considering the Wage Census, require very careful definition, and many subsidiary questions must be put to get a proper estimate; the simple query, "what is your weekly wage or annual income?" would be answered on so many varying principles that the result would be valueless.

Definiteness.

Thirdly, the questions must be such as will be answered truthfully and without bias There is hardly a demand on the census form which would not be excluded, if this rule was too rigorously enforced, as we shall see immediately. The worst offender in this respect is the

Veracity.

* But see p. 138, *infra*.

question, *Employer or employed?* For though there are many cases in which a man is both employer and employed so that this question should be excluded by our second test, many persons consciously exaggerate their social importance by erroneously replying the former. Questions relating to social position must generally be excluded by this rule.

Fourthly, the questions must be those which will be answered willingly, and must therefore not be inquisitorial, or such as to raise apprehension of a change of law or an imposition of taxes. Questions as to membership of trade unions, or of friendly societies, or as to insurance, would be thought inquisitorial. Many would refuse to state their incomes, holding it to be no one's concern but their own. Questions as to rent might be regarded as possibly leading to taxation. Questions as to religion are badly answered, as was shown in the evidence before the Census Committee of 1890,* and should be excluded by each of these four rules. Some persons do not know what their religion should be named, others would find the question indefinite, others would deliberately answer wrongly, and many not at all.

Reluctance to answer.

The questions on the census form † not excluded on one or other of these grounds are Nos. 1, 2, 3, 4, 5, and 10; these are fairly definite, and householders are generally able and willing to give correct answers to them. Questions 6, 7, 8, 9, and 11 compete with many others, which lead to equal inaccuracies, for a space on the census sheet. No. 6 has long held its place because of its great importance; Nos. 7, 8, and 9 are on their trial. A further discussion of the merits of some of these is to be found in the Report of the Committee already mentioned; here it is only intended to indicate the general grounds of inclusion or exclusion.

So far we have not discussed the important question as to who should fill in the form. If, as in the English Census, it is to be filled in by the householder, the questions must be much simpler in matter and words than if it is to be filled in by an official teller. In the latter case the form may be much more complicated, the questions more inquisitorial and such as might lead to indefinite answers on the part of ignorant people; for the teller would insist on

Filling up of the form.

* *Report of Committee on the Census*, 1890 (C.—6071). † Facing p. 23.

an answer, be able to exclude those obviously wrong, and cross-question till the indefinite answers were so altered as to allow definite tabulation. In a great and complex undertaking like the Census, where many tellers must be impressed for a single day's work, their instructions and the general plan must be sufficiently simple; but as the extent of an inquiry contracts, the tellers can receive more complete instructions, and the information requisitioned may be more complex. This is of most importance in connection with columns 6-9.

The general shape and appearance of the sheet needs attention. If the structure of the family is to be shown, the answers are best given on a single sheet, which must contain enough lines for the largest ordinary household, so that the trouble of fastening together of many couples may be avoided, and tabulation not be hindered. The spaces must contain plenty of room for answers in uneducated handwriting, without making the whole so large as not to lie easily on a desk. The instructions must be distinct and visible, and placed in close connection with the answers; to further this, a skilful use may be made of capitals, italics, and different founts of type. On the form facing p. 23, those in use are roughly reproduced in miniature.

Shape of blank form.

The form should always show for what purpose the figures are collected, and how they will be used, in order to enlist the support of the informant and allay misapprehension. The extent to which this should be done depends a good deal on whether the filling-up is compulsory, as in the population census, or voluntary, as in the wage census. In the case before us no preamble is necessary, since every one knows the main features of a census, and most are willing to further its objects; but it must be shown that the inquiry is sanctioned by Parliament, and that compliance is compulsory. This is done on the back, on the fold which is outside before the form is opened; and even though penalties are threatened against absence of or falsification of returns, the last sentence describes the object of the inquiry and guarantees the informant against malicious use of his answers. Where information is voluntary, a careful letter should be printed and circulated with the form, persuading the informant to give his assistance.

Purpose to be shown.

While the main part of the form is filled in by the house-

holder, other parts are filled in by the officials, and with very little trouble a good deal of subsidiary information can be collected in this way. On the outside the Parish, Town, Sanitary District, Street, and Number are endorsed, so that the answers can be tabulated for any of these districts. The teller could also, as he took the form, enter the number of stories to a house, which is not done in the English Census, and other information as to the style of house and street might be endorsed. In a more intensive investigation, Mr Charles Booth's assistants, for instance, could be trusted to come out of a house with an accurate knowledge of many interesting details.

Subsidiary information.

We can now proceed to the individual criticism of the form in the light of the rules suggested above. In the first place, even the arrangement of columns is not perfect. To labourers who are not in the habit of writing at all, and who have (to judge from election posters) to be instructed how to put their mark in the right place on a ballot paper (many papers being destroyed simply through ignorance), this arrangement of horizontal and vertical columns would be confusing, and without help they would not gather at all what they were to do. They would fill up more easily a paper in which the answers were to follow the questions immediately:—

Lines and columns.

State your Name_____

State your Age_____

State your Sex_____

Unmarried, Married, or Widowed_____

and so on.

This form, however, could only be used if a separate paper were to be filled in for every individual, children and all. Other elementary matters might be improved. On looking through the form a great number of words and phrases will be found which are not in common use, *e.g.*, abode, ~~dwelling (as a noun)~~, elsewhere, East Indies, imbecile, "precise" infirmity, general term, column, the foregoing, condition as to marriage. In column 1 the phrase, "name and surname" reads as though surname were not a name, and perhaps the word "surname" is not in general

use, so that the printed word might be taken to mean title, and the confusing answer "none" written under it. Does the instruction "write after" mean to the right, or below?

The first question, which for the general purpose of the census should be the most definite of all, leaves some room for doubt. What of a night-watchman returning at 4 A.M., or a printer at 2 A.M.? What constitutes a traveller: does a man who leaves the house before midnight, or a man who goes down to Brighton by the theatre train come under the term? Is midnight or 2 A.M. the critical time? What of a person who dies at 1 A.M., or a birth at midnight? How is the householder to know whether any of his establishment are returned elsewhere? Since too many instructions only lead to confusion, the tellers should be specially taught the answers to such questions.

Criticism of the questions.

"Slept or abode."

The very meaning of the phrase "population of a district" is open to much doubt. In France "la population de fait," which consists of all present in the given district at the given moment, is distinguished from "la population de droit," which consists of all usually resident in the district, including those temporarily absent, and excluding those only momentarily present, and from "la population municipale," which is "la population de droit," less prisoners, hospital patients, scholars resident in schools, members of convents, the army, and so on.* The English Census counts "la population de fait." In the United States we find a "constitutional population," which excludes residents in Indian Reservations, the Territories, and the District of Columbia; the "general population," which includes in addition the Territories (except the Indian Reservation, Indian Territory, and Alaska); and the "total population," which includes all excluded in the former.† In the future questions will arise as to the inclusion of the Philippines and Cuba. Notice that the Channel Islands and the Isle of Man are included in the English Census.

Meaning of population.

---

* See Bertillon, *ibid.*, p. 146.

† Willcox: *Area and Population of the United States at the XI. Census*, a book which gives a very useful criticism of the accuracy of the most elementary data of statistics. It is a pity that space is wasted in a useless attempt to supplant the word "statistician," which has now a definite meaning, by the word "statist," which has another equally definite meaning. Does Dr Willcox wish to substitute "statics" for "statistics"?

It is possible to find difficulties in filling up all the columns except No. 4. For illustration, consider how column 2 should be filled in in the case of a cousin who was a "paying guest," or a relation who was a visitor; for column 3, is a divorced person single or a widower, and what of a woman who is doubtful whether her husband is lost at sea? Errors come from No. 3 because many unmarried people call themselves married.

Age. It is well known that column 5 is wrongly filled in for two reasons—one, that elderly people often do not know their ages accurately and enter them to the nearest round number, so that the returns congregate at 40, 50, 60: the error thus arising is eliminated by tabulation in the groups 35-45, 45-55 years, &c., and for more minute tabulation the groups 3-7, 8-12, 13-17, &c., are suggested: the other is that many ladies habitually enter their ages too low; in this case also the Registrar-General is able to deduce nearly correct totals.

It is to be noticed that, since the ages stated are those "last birthday," the age will on the average be given six months too low, and, in fact, the ages given as 17, *e.g.*, should be scattered nearly uniformly over the months to the eighteenth year.

Occupation. The most important criticisms of the census-schedule are to be made on columns 6-9. It will not be expedient here to go into all the questions raised before the Committee on the Census as regards an industrial census. While there can be little doubt that a thorough census of occupations would be best undertaken separately, and on somewhat different principles from the population census, it is certainly better, till opinion is ripe for so radical a change, to include in the present census the best questions we can as to occupations, than to omit them altogether in despair of accurate results.

The objects aimed at, which we must always keep in mind when criticising special questions, are two: to find the number employed in each trade and industry, that is, so to say, to form vertical divisions; and to find the number in each rank or grade of employment (labourer, artisan, employer, &c.) in horizontal divisions; so that the tabulation may give some such result as—

TEXTILE INDUSTRIES.

	Cotton.	Wool.	Linen.	Totals.
Employers -				
Managers -				
Overlookers -				
Spinners -				
Weavers -				
Labourers -				
Children -				
Totals -				

The necessary minimum of information would be given by such answers as

Legal—Solicitor—Managing clerk.
Mining—Coal—Hewer.
Metal-worker—Iron—Smith's striker.

Now the simple instruction, "State your occupation," would of course not lead to information of this sort. The coal-hewer would simply say miner; the clerk, managing clerk; the striker, very likely smith. To explain what is wanted and avoid mistakes, the question is not put on the face of the form at all, but the informant is referred to the back, half of which is devoted to instructions relating to this column. These are lucid, carefully picked out with capitals and italics, comprehensive, brief and to the point. No one who wishes to fill in the form rightly, and is sufficiently educated to understand simple instructions, can easily go wrong. Yet, as a matter of fact, these instructions are in very many cases neither read nor followed; and this fact is very important in connection with the general study of blank forms of inquiry. Forms issued to people uninterested in the object in view will generally be filled in with the least possible expenditure of time and intelligence. Hence two courses are open: to reduce

the question to the simplest possible form, and make the best of the result; or not to allow the informants to write in their own answers, but to take them *vivâ voce* by means of a teller, who has mastered the instructions, and has the necessary legal force behind him to compel information. The latter course entails time and expense.

The result of the present system of inquiry, combined with a faulty method of tabulation, which it to some extent makes necessary, is that we have no reliable census of occupations for the United Kingdom. The present figures break down both from faulty data and from insufficient tabulation directly we attempt to make any calculations depending on them.

The result of the new questions.

An attempt has been made to correct to some extent our ignorance of the relative numbers of unskilled and skilled labourers, employers and employed, by columns 7, 8, and 9. The headings are not a model of clearness; there is not the ordinary imperative "state" or "write," nor is one told on the front of the form whether to write Yes or No or to make a mark in the appropriate column, nor is the distinction between the three headings a perfectly definite one; but still one is hardly prepared for the following statement in the report: *—

"In numerous instances, no cross at all was made; in many others, crosses were made in two or even all three columns, and, even when only one cross was made, there were often very strong reasons for believing that it has been made in the wrong column. Oftentimes this use of the wrong column can scarcely have been other than intentional; being dictated by the foolish but very common desire of persons to magnify the importance of their occupational condition. This desire must have led many subordinates to return themselves as employers rather than as employed, for it is only on this supposition that we can account for the otherwise unintelligible fact that, under several headings, there are actually, according to the returns, more employers than employed, more masters than men. . . . We hold [these returns] to be excessively untrustworthy, and shall make no use whatsoever of them in our remarks."

This attempt and its result are of the greatest importance to all who try to draw up forms of inquiry.

---

* *General Report on the Census of* 1891, p. 36 (C.—7222 of 1893).

Before leaving the subject, it should be mentioned in passing that we cannot deduce directly from our census the number of persons dependent on a particular trade for their living; that is to say, the number of employers, their families (not otherwise returned) and domestic servants, and the number of employés and their dependent families. This, the most important total for estimating the relative importance of different trades of the country, is not tabulated, though such tabulation has been found possible in other countries, and we are dependent on the estimates of statisticians for such totals.*

To see how the information given by the answers on the census schedule can be worked up into detailed specific numbers, it is only necessary to look at the diagram and table prefixed to each of the sections relating to special trades in Mr Booth's *Life and Labour of the People* (*e.g.*, vol. v., p. 46). †

* See Booth in *Statistical Journal*, vol. xlix. † See p. 78, *infra*.

SECTION 2.—THE WAGE CENSUS.

The main differences between the wage census, taken in 1886, and the general population census are—(1) That the filling up the forms in the wage census was voluntary; (2) that their correct filling up required a higher degree of intelligence and education. As before, we must consider first the object which the wage census was intended to fulfil: it was to describe the earnings of the people of the United Kingdom, to compare the rates of wages trade by trade, and to find the relative numbers earning at each rate. What is the best quantity to measure with this object in view? As a preliminary question should we take the day, week, or year as the unit of time? Clearly we shall not be able to compute weekly wages if we only obtain daily, for the week's work varies from four to seven days in different occupations. The week's wage is a more definite quantity; but the simple comparison of weekly wages in different trades will be deceptive, because most trades are busier at one season of the year than at another, and in many the difference between season and season is very great; in any particular week, then, we may be comparing the best season of one industry with the worst of another. To avoid this error, and because we do not know how many full weeks' wages are obtained in a year, except in a few non-intermittent trades, it would seem best to take the year as unit; but the direct calculation of an individual's annual earnings is practically impossible. The employer is not acquainted with this sum, for in large establishments the hands are continually changing, and one man will be paid by two or more masters in the same year; and even in a factory with a nearly constant personnel, the weekly amounts paid to individuals are not in general so tabulated as to be easily summed, and the working out of the totals would require a prohibitive amount of clerical labour. If we turn to the workman, on the other hand, we shall find in the majority of cases that no accurate account has been kept of earnings through the year, and it would only be by careful individual examination, impracticable on any large scale, that

The object.

The unit of time.

an estimate could be made; in many cases the men, even if willing, would be quite unable to give a connected account of their earnings during the past twelve months.

It seems clear that we must adopt a smaller unit, and since most wages are paid weekly, a week is the most natural one. The subsidiary questions which will lead best to an estimate of annual earnings will be discussed below. The answer to the former question, as to the best quantity to investigate, is indirect; the only individual measurements we can obtain directly are the week's wages, but these may be supplemented by estimates *en masse*.

Next, who possess the information we require? Clearly both employers and employed, and in an ideal census the answers would be obtained from both groups; but considerations of simplicity, cheapness, and accuracy are all in favour of applying to employers alone.

Employers and employed as informants.

If employés were to be interrogated the procedure would be as follows. Draw up a form on the analogy of the census form, describe very briefly the purpose of inquiry, add a short series of concise, lucid, simple questions in suitable type and with careful spacing, such as will lead to the minimum information required; let these forms be left to be called for, and when collected, let the tellers have time and opportunity to examine and correct them. It is clear that this method would entail an even more expensive organization than the population census, and as the result of experiment it may be doubted whether the maximum of accurate information that could be thus obtained would come up to the minimum that would be of use. A partial inquiry could, however, be carried out by means of trade unions if they were willing to give serious assistance.

The method of inquiry among employers was as follows:— Suitable blank forms and an explanatory letter were sent by post to all employers, whose addresses could be found, in the industries selected for investigation, and the answers were returned to the central office by post. This is far simpler and cheaper than the suggested scheme for inquiry among workmen, requiring far fewer forms and only a small staff of clerks. With business men it is a simpler matter to post the return when completed than to keep it for collection by hand. Since there is no personal intercourse over the matter it is especially necessary that the questions should be

lucid, for the additional correspondence necessary to rectify errors is a source of worry at both ends. A copy of one of these forms, abridged only in the number of subdivisions, is subjoined here and on the following page.

WAGE CENSUS.

RETURN OF THE RATES OF WAGES PAID IN SILK MANUFACTURES.

*Name of Factory or Firm*_____

*Address*_____

_____

Note.—It is requested that the salaries of clerks and managers may be excluded. The return is of wages of working men only.

Numbers employed on_____ 1886 - - No._____

Amount paid in Wages in the year 1885 - - £_____

Highest weekly amount paid in 1885 £_____ Date_____

Number of Hands paid in that week - - No._____

Lowest weekly amount paid in 1885 £_____ Date_____

Number of Hands paid in that week - - No._____

State the present average rate of pay for overtime: that is, whether overtime is reckoned as time and a quarter or time and a half, &c., or in what way reckoned_____

State whether overtime is at present being worked, and how much; or whether less than full time, and how much less_____

CURRENT RATES OF WAGES AND HOURS OF LABOUR PER WEEK of Persons employed in each Branch of the Silk Manufactures, on_____ 1886.

DESCRIPTION OF OCCUPATION. *N.B.*—It is requested that this list of occupations may be revised where necessary.	CURRENT RATES OF WAGES PAID and NUMBER OF HOURS OF LABOUR PER WEEK when in full work, but exclusive of Overtime. *Note.*—State the Number of Hours of Labour per Week, whether the Workers were paid by Time or Piece-work, and if paid by Piece-work give the amount earned in a week, exclusive of Overtime.											
	MALES.						FEMALES.					
	MEN.			LADS & BOYS.			WOMEN. 18 years and upwards.			GIRLS. Under 18 years.		
	Number Employed.	Rates of Wages.	Hours of Labour.	Number Employed.	Rates of Wages.	Hours of Labour.	Number Employed.	Rates of Wages.	Hours of Labour.	Number Employed.	Rates of Wages.	Hours of Labour.
Silk Throwing—												
Parters - { Time / Piece												
Winders - { Time / Piece												
Cleaners - { Time / Piece												
Spinners - { Time / Piece												
Doublers - { Time / Piece												
&c.												
Silk Spinning—												
Openers and Sorters - { Time / Piece												
Boilers - { Time / Piece												
Dressers - { Time / Piece												
Preparers and Carders - { Time / Piece												
&c.												
Silk Weaving—												
Winders - { Time / Piece												
Warpers - { Time / Piece												
Warp Pickers or Clearers { Time / Piece												
Doublers - { Time / Piece												
Fillers - { Time / Piece												
&c.												

The measurement of the annual earnings of groups of workpeople was the ultimate object of the inquiry. Annual earnings are composed of many different items, of which the following are the most important:— Ordinary weekly wages, pay for overtime, special payment for special work (*e.g.*, of builders if sent to a distance), or at special seasons (such as the harvest); and payments not in cash, such as free or reduced house-rent, free or cheap coal, and special goods at cheap or wholesale prices (such as cloth in textile factories, or potatoes for agricultural labourers).

Annual earnings.

When payment in kind is at all general or important, it is generally better to proceed on a different method entirely, *e.g.*, that followed by the Agricultural Sub-Commissioners of the Labour Commission. When it consists of only one simple item, such as a house rent-free, it can form the subject of an additional question on a form similar to that on p. 35. In the silk industry this does not occur; but this discussion shows the necessity of preliminary knowledge on the part of the investigator before the right form of inquiry can be drawn up.

We have left for consideration the weekly wage, and overtime and special payments, the last two of which can be grouped together. The ordinary weekly wage is a sufficiently general and definable quantity in most subdivisions of most industries. A foreman could generally state how much is earned in an ordinary full week for each of the hands under him. In many cases there is an hourly or weekly sum regulated by a trade union, as in the building trades. In others, as in the cotton industry, piece-rates are so regulated as to bring out a definite sum for the week's work graduated in relation to the difficulty of the task; in general, a very rapid survey of the wage-book will show what the worker in each subdivision will make on an average. Thus the average weekly wage in an ordinary full week can be found with considerable accuracy, but this takes us only part of the way in the calculation of annual earnings; we need to know in addition to this how many full weeks are made in the year. It is the method by which this is attempted on the printed form that is open to most criticism. The questions used are on page 35, and afford a good example of the general difference between the quæsita and the data which are attainable. The quæsitum is: To how many full weeks' wage are the annual earnings equivalent, allowing for slack weeks and overtime?

The first crucial question to decide is: Are we to allow for an average loss of time, say a week in the year, through sickness, or are we to allow only for time lost through failure of work? Since sickness is an individual not a general misfortune, it will be better to exclude it if possible. Now overtime in one season, especially if its wages are on "time-and-a-quarter" or "time-and-a-half" basis, very quickly tends to balance slack time at another season, though it may be supposed that it is rarely the case that more than the normal week's wage is averaged through the year. Thus it will be logical as well as simple to estimate the year's earnings as so many normal weeks' wages. For example, if we found that two weeks were lost through sickness and three through the mill stopping, and that overtime in one busy month had added wages equivalent to two normal weeks, we should have forty-nine weeks' full wage. The figures which will give this result will be the total sum paid in wages in the factory in the year divided by the aggregate normal week's wage of the people dependent on the factory, supposed all at work. Thus, if 1,200 hands (men, women, and children) would, if all at work, make £1,000 in a normal week, and this was the average number dependent on the particular mill, and if £48,000 was paid in the year in wages, annual earnings would be equivalent to forty-eight normal weeks, and earnings would average £40. Now the total paid in wages is generally kept separate in business accounts, but the number dependent on the mill for work is often not known accurately; for the personnel of a large establishment is subject to continual change, and the manager would not know whether a person who left went to another mill or got no work. The total number of all who had worked there during the year would be too great for this purpose, and the number at work in a normal week too small. The number open, perhaps, to least objection is the number at work in the busiest week of the year; for those absent except through sickness when trade is busy cannot be said to be dependent on the factory, but if not at work elsewhere are among the permanent unemployed; very few workpeople indeed will be taking their holiday at a busy time, and it may reasonably be supposed that all the factories in the same industry will have their busy and slack seasons at nearly the same time. The answers then to the printed questions—Total paid in year, and number of hands in busiest week—tell us all we need to know,

Quæsita and data.

if we may make this assumption; for then the total sum paid as wages in the year, divided by the maximum number employed in the busiest week, gives the average annual earnings. To find the equivalent number of normal weeks, multiply the maximum number employed by the average wage found on the second page of the form, so that the product shows the aggregate weekly wage if all were employed, and divide the total paid in the year by this product.

Lost time in the cotton industry.

In the Cotton industry the sum of the greatest numbers employed (if these may be taken as equivalent to the numbers employed in those weeks when the wage bill was highest) was, in 1885, 87,887. £3,148,566 was paid in wages in that year in the factories making returns. Average annual earnings were therefore $\frac{£3,148,566}{87,887} = £35.\ 7s.$ The average wage in a normal week in 1886 was 15s. 2½d.; the product of this and 87,887 is £66,830. The equivalent number of normal weeks' work is $\frac{3,148,550}{66,830} = 47.$ Hence we may conclude that, if our basis of calculation is correct, five weeks was the average lost time at that date.

This is not the method adopted in the General Report of the Wage Census; there the total paid in 1885 is divided by the number employed in a given week in 1886. This number is certainly too small, less than the number dependent on the trade, and as might be expected gives on analysis absurd results in some cases. It is to be noticed that the method here described cannot be employed in those few industries which the employés are able to leave in the slack season in order to earn wages in other trades which may then be exceptionally busy.

Since there is no reason why the number absent through sickness should differ in the busiest week from the average number so absent, it is clear that the estimate we obtain for average lost time (five weeks in the wool industry) is in addition to the average time lost through sickness; this may often be estimated from the returns of friendly societies.

In the corresponding French wage census, of which the results were published in 1898,* an estimate of the number of days' work obtained in the year is formed on a different basis.

* *Salaires et Durée du Travail*, 1897, pp. 15, 16.

The data collected were—(1) The variation each month of the personnel in each industry, which is found to average 4 per cent. for the year—that is, for each 100 employed, 96 are found who have been in the same establishment for as much as twelve months: (2) The differences between the maximum and minimum numbers employed in each establishment month by month during the course of a year, which are found to average 19 per cent. of the (? average) personnel. From this we may perhaps draw the conclusion that, on an average, half this number, at least, are in general out of work: (3) The number of different persons who have been employed in each establishment at one time or other in the year; this is found to be 140 for each 100 permanently employed, from which the legitimate conclusion is that the average number of unemployed is not so much as 40 in 140, *i.e.*, 28 per cent. These two percentages, 9 per cent. and 28 per cent., are taken to be the inferior and superior limits of average lack of work. This information is more detailed and perhaps more reliable than that on which the method, used above for the English figures, is based. Data obtained from syndicates of French workmen indicate about 20 per cent. as the average want of work; the English figures obtained by the method described above from the whole wage census yield about 12 per cent.

The French method.

This somewhat lengthy discussion on the few questions included on the first page of the form is a good illustration of the necessity of considerable preliminary study before a blank form can properly be drawn up. Space does not allow a detailed criticism of the rest of the form.

SECTION 3.—THE WORK OF THE LABOUR DEPARTMENT.

The Labour Department of the Board of Trade was founded in 1893; its functions are to collect and publish information, chiefly statistical, relating to the economic conditions of workpeople, and the state of the market for labour. Its work lay almost entirely in virgin ground; new sources of information had to be tapped, new methods developed. While it was untrammelled by tradition, it could avail itself of the experience of the Board of Trade, and was already in touch with a widely extended organization. Under these circumstances it was soon able to attract a comprehensive and continuous supply of valuable information; and the methods by which it accumulates and compiles its statistics should be interesting and instructive to all those whose business it is to work in any statistical field.

The Labour Department.

The figures which are received periodically by the Department are published monthly in the *Labour Gazette*. Here the first article each month is on the "State of Employment." As before we must first consider the question, What are the exact objects of the investigation of which the results are here published? They are to find out how many persons are out of work in each trade and district, what percentage they form of all dependent on each industry, and how this percentage changes month by month and year by year. The next question is, How much of this can be discovered, and, if we cannot measure these numbers directly, what are the best allied quantities to measure? Since no universal register is kept of the unemployed, it would seem easier to estimate the number employed, since an employer can generally state how many workpeople he has at work at any given time. If we cannot discover the number of men at work, we may perhaps be able to find the number of machines, furnaces, mines, &c., at work, and deduce the number of men employed with them, and thence the number of unemployed; or we may find for how many hours work was carried on in a factory or a mine; or we may even go a step further back and find the amount of goods produced, and thence

Unemployment.

Possible measurements.

estimate the other quantities; or we may learn the total amount paid in wages. These are the numerical methods; but there are others, useful if not so exact. We can obtain reports as to the condition of employment in the various districts or industries, not in numbers, as is generally necessary, but with descriptive adjectives,—such as busy, slack, improving, much the same,—which may lead to numerical estimates, or may serve to check results. Lastly, organizations for facilitating employment may send in returns of the applications made to them. Nearly all these methods are in use at the Labour Department.

Informants.

Next, who possesses the necessary information? As regards the number unemployed, the only registers kept are those of trade unions, to whose secretaries inquiries should be addressed. The figures so obtained will naturally only relate to those sections of an industry where trade unionism exists. As regards the number employed, the masters are the authorities, and forms must be sent to them asking the numbers at work day by day, or at longer intervals. With respect to the number of machines at work, the number of shifts, and the total wages paid, the masters again have the information. For the amount produced, the masters, or in some cases officials to whom they make returns, can supply the facts. For general information as to the state of employment, some presumably competent person, in touch with all the factories in an industry, or all the trades of a district, must be impressed to forward periodical reports. The Labour Department is in touch with a great number of such correspondents, many of them connected with the trades councils of their towns.

The question as to whether the information will be given impartially and willingly need not detain us long in this case; for, generally speaking, the returns are simply automatic copies or registers of known numbers, and would only be partial if wilfully falsified; and since the returns are made periodically, the persons concerned regard them as a matter of course, and, once they have commenced, continue willingly to forward the requisite figures.

By the courtesy of the Labour Department I have been able to obtain copies of most of the forms in use. There are some forty in all, each suited to some special industry or method of investigation.

It must be remembered that the Labour Department had to

form its own intelligence organization, and initially was obliged to apply to persons able to give information, just as any private investigator would. A connection had therefore to be established with trade unions and other societies, and with manufacturers; and, when a nucleus had been formed, continual efforts were necessary to extend the organization in all directions. One or two of the circular letters written for this purpose are given on the following pages, since they are typical of the method which investigators must employ to enlist the help of possible informants who are uninterested. The points to notice are:—(1) The statement of the exact purpose for which the information is wanted; (2) the simple and explicit direction as to what is to be done by the informant; (3) the undertaking that the information will not be used in any way that can do, or appear to do, him injury.

The formation of an intelligence department.

Here is one of the earlier letters, opening a connection:—

LABOUR DEPARTMENT, 1894.

DEAR SIR,—The Labour Department of the Board of Trade, which is charged with the duty of collecting periodical statistics as to the condition of the Labour Market, is desirous of obtaining fuller information from month to month with regard to the state of employment in the Pig Iron Industry. For this purpose, the Department would be glad to receive monthly information from a large number of the employers in the United Kingdom as to the number of furnaces in blast and the numbers of workpeople employed, on the average, at each furnace.

I shall accordingly be glad if you will be kind enough to assist the Department in making this inquiry complete by filling up and returning to me *before the 4th of May* the enclosed form. Postage need not be prepaid if the reply is addressed to "The Commissioner for Labour" at the address given above.

The results of the inquiry will not be published in such a form as to render possible the identification of particular returns.—Yours, &c.

When the Department had organized its work, and tabulated and published some of its returns, the next step was to endeavour to achieve completeness. When many are known to have given information, the more cautious will be encouraged, the less energetic be ashamed to be less public-spirited than their neighbours, and the critical anxious to correct mistakes. The first of the following letters, which is used for general purposes, takes ad-

vantage of these tendencies, and the second is another excellent example of the method of extending the organization :—

1895.

DEAR SIR,—I am forwarding herewith a copy of the "Labour Gazette" for the current month, and beg leave to draw your attention to the article therein dealing with the state of employment in the _____ The Labour Department is very desirous of making the information contained in these monthly reports as complete as possible, and trusts you will kindly assist by filling up and returning the enclosed form. You will notice that the form is of a very simple kind, and one that can readily be filled up without much trouble.

I may add that Returns are regarded as strictly confidential and are only used to produce general statistical results in which the identity of individual returns is lost.—I am, &c.

March 1895.

DEAR SIR,—This Department has for some time past received monthly Returns, both from the Dock Companies, and the Ship-owners who do their own unloading work in the port of London, with regard to the number of Dock Labourers employed. These Returns are collected with a view to throwing light on the periodical fluctuations in the employment of this class of labour; but the figures are published in a general total and not in such a way as to make possible the identification of particular firms supplying the information. The article on page 36 of the enclosed Labour Gazette will show you the use made of the Returns.

Hitherto no exact information has been obtained with regard to employment of labour at the wharves, and you will readily see that the addition of such information would very greatly increase the value of the statistics. The managers of several of the most important wharves on both sides of the river have been good enough to promise to make monthly Returns; and I should be greatly obliged if you could see your way to assist the Department by supplying the information specified on the enclosed form, not later than the date there indicated.

You will observe that a form is provided for the daily number of labourers employed, and, alternatively, for the average weekly number. The daily number would, on the whole, be the most useful for the purposes of the Department; but if for any reason you cannot see your way to supply such detailed information, a weekly average would be of value.—I am, &c.

Another letter may be given as serving to encourage those who have already engaged in the good work. It will be noticed that, though now more concise, it is still insinuating.

Agricultural Labour in January.

DEAR SIR,—I am instructed by the Commissioner for Labour to ask you to be good enough to favour this Department with replies to the questions on pages 2-4 of this form, by Friday, 4th February 1898.

I beg at the same time to thank you for your kindness in sending answers to questions put to you by the Department on former occasions.—Yours, &c.

These letters are well worth noticing because they have assisted to build up a very efficient organization for information out of nothing, and have succeeded in eliciting answers from uninterested men of business, who are not given to spending time and trouble on unremunerative labour.

Please forward this Return to the address on the back not later than the Fifth *of the month succeeding that to which it relates. No postage need be paid.*

RETURN OF STATE OF EMPLOYMENT

*in Month of*_____189

Name of Society_____

Total number of members in Society at close of month_____

Number receiving out-of-work pay in last week of month_____

(Do not include members on strike or locked out.)

State, if possible, number of members entirely unemployed but not receiving benefit in last week of month_____

State of employment for month_____

If any dispute, change in wages or hours of labour has occurred, please say, and the necessary forms will be forwarded at once_____

Remarks_____

Signed_____

Secretary.

Date_____189

The form given above is that issued to trade-union secretaries, and it is by its means that the only perfectly definite measure of want of employment is obtained. It should be remembered that it is filled in monthly by the same official, and requires no special explanation. Since the first attempt to draw up such a form is apt to present difficulties, this may be noticed in detail. First, we find an instruction as to the way it should be returned. Most forms are provided with a printed envelope to save trouble and mistakes, and postage is paid by the investigator. Next comes a brief heading and the date, and then the name of the society, an item used for reference and further inquiry, but not for publication. Now we need to know chiefly the percentage unemployed, but secondarily the total number. The questions most easily answered should be asked, and the calculations done at the central office, for the trade-union secretary may make arithmetical mistakes. Again, it is not the numbers day by day that are asked, for they are hardly known; nor week by week, which would give trouble; nor even the average for the month, for that might lead to guess-work; but a definite day or week is decided on, the same for all trades. For purposes of comparison trade with trade, or month with month, this is found sufficient.

Blank forms.

Examination of unemployment figures.

We notice next an important point connected with the definition of Unemployment, and also an illustration of the necessity of studying the figures at their source. It is not stated explicitly in each *Gazette* whether men on strike or locked out are included as "unemployed" or not. A reference to this form shows that they are not included, and, therefore, before conclusions are drawn as to the amount of work obtained year by year, the excellent statistics relating to labour disputes given in another part of the *Gazette* must be studied.

Definition of unemployment.

All members of trade unions out of work do not at once receive "benefit," the technical term for any payment of union funds, and therefore a correction must be made in some cases for those who have not yet come "on the society." The number is likely to be known accurately to the local secretary; but if the form had to be filled in at a London office, say, for the whole Amalgamated Society of Engineers, the number would not there be known. At this point we are left in some

doubt as to the methods of the Department; for we are not told whether these forms are sent to all local secretaries, or whether they are filled up centrally for whole districts, or what additional information is obtained from other sources.

The next line is for a qualifying adjective which will serve to check labour correspondents' information, and to indicate whether the last week was typical of the month.

When an organisation is ready for a special purpose, any secondary use may be made of it that will not vitiate its chief end. The Labour Department is always anxious to hear of all changes of wages, and in general has to detect their existence for itself; hence no opportunity of obtaining such information is lost, and this widely circulating form is used for the purpose.

Subsidiary information.

Lastly, if the informant is an intelligent man and acquainted with the methods of the Department, it is well to give him an opportunity of adding any relevant remarks that may occur to him. On the line "Remarks" might be given some reason for any exceptional numbers or information as to trade prospects which might furnish a clue to the Department for other investigations. The paper is signed and dated to show that it has been filled in officially and at the right time.

These forms are not always filled in and returned to date; sometimes special application has to be made for them, and occasionally the necessary numbers have to be interpolated from other sources.

The next form given is more complicated, and illustrates two of the methods mentioned, finding the number of persons employed and the number of days worked.

Please fold and return this form by the 29th December *to the address given on back.*

EMPLOYMENT AT COAL MINES.

Individual Returns are regarded as confidential and not published separately.

County in which Pits are situated_____

Name of Firm or Company_____

Postal Address to which form should be sent_____

Number of shifts usually worked in each 24 hours_____

Names of Pits or Seams.	State whether *bulk* of Coal raised was "House," "Steam," "Gas," "Manufacturing," or "Coking" Coal.	* Number of Workpeople paid on last pay-day in four weeks ending		† Number of Days on which Coal was hewn and wound at the Colliery in four weeks ending December 25, 1897.	‡ Short Time. If any of the days stated in Col. 5 were shorter than usual, please state in Col. 6 the total amount of time to be deducted by Labour Dept. in four weeks.	
		25th Dec. 1897.	26th Dec. 1896.			
1.	2.	3.	4.	5.	6.	
		No.	No.	Days.	Days.	Hours.
No. of "Other Workpeople" general to all or several Pits and not included above, and Number of Days worked by them in four weeks						

* The *number of Workpeople* should include all Men and Boys, &c., employed in and about the Pits, except Clerks and Managers. The number should also include "Drawers" and others who may be paid by "Hewers."

† The *number of Days*, whether full or not, on which Coal was "hewn and wound" should be inserted in this column. If on any of these days short shifts only were worked, the extent of the time lost should be stated in Column 6; but it should be left to the Labour Department to deduct from the figures in Column 5 the Short Time, if any, given in Column 6. If the time worked on Saturday is usually shorter than on other days, no reduction should be made on that account.

‡ *Short Time.*—Please state here any special reasons for Short Time :—

Note.—A General Summary of the Returns received will appear on 15th January in the *Labour Gazette*, which can be ordered through any Newsagent, price 1d.

The form has one or two peculiarities. A colliery company has often several pits, so that in the first place it is not obvious at which address this information can be most readily given, while it is important not to waste time and trouble in forwarding from office to office; and, in the second place, not only will work be done for different lengths of time in different pits, but also there will be variation from seam to seam in the same pit. This was not recognised in the first form sent out, but a second had to be sent distinguishing the pits and asking for subsidiary information. In this form may be seen the modifications that must be introduced to suit the questions to particular industries. In a colliery the factor which determines the state of employment is generally not the number employed, but the number of days' work, the number of days "coal is wound." A colliery at full work may make four, five, or six days a week, or eleven days a fortnight (leaving one day a fortnight for repairs), according to the custom of the district and the state of the trade, and there may be two shifts or three in the twenty-four hours. If work is slack, the number of shifts per fortnight, which is really the essential quantity to know, will be diminished, and the alteration will very likely affect all employés equally. Again, since the colliers are not all at work at once, the question is not "how many are at work?" but "how many are paid?" the pay-day, once a fortnight, or however often it may be, being perhaps the only time when all the workers are together. The number at pay-day is, therefore, the number employed in the mine, a quantity varying as new seams are opened or old seams worked out, and the number of days on which there is work is the factor which determines the amount of work obtained per workman. Notice that the questions, number of shifts per day, number at pay-day, and days at work, are precisely those which the manager will find easiest to answer. Since, however, days are of different lengths, depending on the demand for coal, the good working of machinery, the presence of the necessary trucks, and the efficacy of the railway arrangements for clearing the yards, and other circumstances, it is necessary to know whether on any working days, winding stopped early or the shift was shortened; hence the question in column 6, which will give the manager more trouble.

Peculiarities of form.

In the form relating to dock labour, the question is simply, how many are employed, not at the end of the month, but

day by day; for labour at the dock fluctuates violently and continually, as may be seen from the monthly diagram in the *Labour Gazette*. On that relating to the Surrey Commercial Dock, the question is again modified to suit, it may be supposed, a special method of bookkeeping, and reads, "What is the amount of wages paid at the end of each week?" Wages are perhaps a better measure of dock labour than number employed since the number of hours worked varies continually, men being taken on for long or short hours, but the rate of pay varies little. On both forms there is a question as to any special holidays or other events affecting work. On the form sent to pig-iron works, the question asked is as to how many furnaces are "in blast," or have been "blown out," or re-lit; on that relating to steel, iron, and tinplate works, the information required is "the number of shifts" worked in four weeks. Another form is to be filled up by a single correspondent for a wide district, and the returns are entered under the headings—Number of mills (1) running full time and giving full employment; (2) running full time, but giving only partial employment; (3) running short time; (4) stopped.

Forms for other industries.

Another instance of adaptation of the form to a particular industry is afforded by the inquiries as to agriculture. In this case the number of employers is very great, they are very much scattered, and little used to statistical inquiries, and the labourers are for the most part uncombined. On the other hand, in the majority of villages agriculture is the predominant industry, every one knows all about every one else and any one intelligent person can give an accurate account o the state of labour in his district. It is necessary then to arrange with a labour correspondent, a farmer, or a member of the Village Council, or the chairman of the District Council, and to apply to him monthly for information.

Employment in agriculture.

Only one general organization is necessary for the collection of the three groups of figures wanted by the Department for all industries. These groups relate to the state of employment, which fluctuates continually; changes of wages, which in some cases take place at stated times, in others occur irregularly; and strikes, which may begin at any time and last a long while. In the case of agriculture, the three groups of questions are placed on a single form, though the practice has changed a good deal since 1893.

One form, that in use in 1894, asks for complete details as to wages and the number employed at the harvest, with a page devoted to strikes, and two spaces for remarks on the weather and on things in general.

The next, July 1895, deals with haymaking, strikes, and wages. The questions here are as follows:—

1. Were there any able-bodied agricultural labourers in irregular work in your Parish during the month of June?

2. If you answer question 1 in the affirmative, can you give the numbers and state about what proportion those in irregular work were of the total number of able-bodied agricultural labourers?

3. If you can give the particulars asked for in questions 1 and 2 for any neighbouring Parishes, kindly do so.

4. What daily or weekly wages are being paid in the district to the regular farm hands during haymaking? Also state how much is paid for overtime and what perquisites are given, such as food, beer, &c.

5. What daily or weekly wages are being paid in the district to extra hands during haymaking? Also state how much is paid for overtime and what perquisites are given, such as food.

6. Were there any agricultural strikes in your neighbourhood during June? If so, please give the following particulars with reference to each strike:—

   (1) The date; (2) The cause; (3) The duration; (4) The result; (5) The number of men affected.

There are differences in the forms for nearly every month in the year; and the questions have been modified as experience suggested till they are finally as follows:—

*County*_____ *Union*_____ *Parish*_____

I.—STATE OF EMPLOYMENT.

1. Approximate number of able-bodied agricultural labourers in Parish.

(If this question has been recently answered by you, you need not repeat your reply.)

2. Were there any able-bodied agricultural labourers in irregular work in your Parish during the month of January 1898?
3. If so, can you say about how many were in irregular work in the *last week* of January 1898?
4. Was employment more regular in January 1898 than in January 1897?

If you can give the above particulars for any neighbouring Parishes, or for the whole of the Poor Law Union, kindly do so.

II.—CHANGES IN RATE OF WAGES IN JANUARY 1898.

Changes in Weekly Cash Rates of Wages of Ordinary Labourers in January 1898.

(*N.B.*—Ordinary labourers do not include foremen, shepherds, cattlemen, carters, waggoners, teamsters.)

Locality in which Change took place. (State whether Change applies to the whole County, or to which Poor Law Unions or Parishes within it.)	Approximate Number of Labourers who have had a Rise or Fall in Wages in January 1898.	Cash Rates of Wages per Week.		Please state in this column for comparison what the Rate of Wages was in January 1897.
		Before Change.	After Change.	
		s. d.	*s. d.*	

*Name of Correspondent*_____

*Postal Address*_____

For convenience of printing the exact spacing allotted for answers has not been introduced in these reprints. In the agricultural forms a great many square inches are allowed for such an answer as "Yes"; in the others the space is allotted in proportion to the amount of information expected. The questions as to wages on this form will be alluded to presently.

Trade disputes.

The information collected by the Department as to trade disputes is detailed and important. The principal questions in the investigation relate to the causes of disputes, the methods of settling them, and the total loss of money to workpeople and employers. Of these the first two are not statistical questions, but are inserted because the inquiry has three objects:—(1) A general examination of the causes of and remedies for strikes; (2) an inquiry as to the course of each particular dispute so as to bring the Conciliation Act into operation if possible, or by disseminating information to assist an arrangement or compromise; (3) the collection of statistical information.

The Department is dependent on its own alertness for knowledge of the existence of disputes, and its chief sources of information are the daily press (London, Provincial, and Trade) and special local correspondents, who are expected to inform the Department, directly work is stopped owing to a dispute, on a special form.

As to the question, Who knows the facts? obviously the only people are the employers and employed; and since they may take different views on all subjects connected with the dispute, both parties must be addressed. In this investigation partiality and bias in the answers will be at a maximum; the questions must be restricted as far as possible to facts about which two opinions are nearly impossible, and any questions which will not be answered willingly should be omitted.

On pages 54, 55 is given the form sent to Trade Unions in 1895, on pages 56, 57 that used in 1897, and on page 58 the letter accompanying the latter.

LABOUR STATISTICS.—
QUESTIONS AS TO

QUESTIONS.	ANSWERS.
1. Name of employer, firm or company, and trade - [Where more firms are involved than one, or the strike or lock-out has been general over a locality, the number of employers or firms to be stated as nearly as possible.]	1.
2. Cause or object of strike or lock-out - - -	2.
3. Whether strike was ordered or approved by trade union - - - - - - - - -	3.
4. Date of commencement and termination of strike or lock-out - - - - - - -	4. Date of commencement.
5. Result of strike or lock-out - - - - - [If dispute has been respecting increase or reduction of wages or hours of work, state exact amount of increase or reduction (if any).]	5.
6. Mode of settlement - - - - - - -	6.
7. Number of persons affected - - - - -	7.
(1) Number directly on strike or locked out	
a. At beginning of dispute - -	
b. At end of dispute - - - -	
[Distinguish between adult men and women, and apprentices or other young persons.]	
(2) Number employed in factories or works where strike or lock-out occurred, and who were thrown out of work thereby, but were not directly on strike or locked out - -	
8. Estimated total amount of wages earned in a full week (exclusive of overtime) by those affected immediately before and after strike or lock-out	8. Before Strike or Lock-out.
a. Directly affected - - - - -	*a.*
b. Indirectly affected - - - - -	*b.*
9. Number of those on strike or locked out who belong to trade unions - - - - - -	9.
10. Amount expended in support of persons on strike or locked out - - - - - -	10. Amount per head per week.
a. By union - - - - - -	
b. By other strike fund - - -	
11. Number of persons who "went in" or returned to work before termination of the dispute -	11.
12. Please suggest means of settling or preventing labour disputes - - - - - - -	12.

STRIKES AND LOCK-OUTS.

STRIKES AND LOCK-OUTS. 1895.

ANSWERS.	GENERAL OBSERVATIONS.
	1.
	2.
	3.
Date of termination.	4.
	5.
	6.
	7.
After Strike or Lock-out.	8.
	9.
Total amount expended.	10.
	11.
	12.

Information for the use of the Labour Department, Board of Trade, 44 Parliament St., S. W.

STRIKES AND LOCK-OUTS.

PART I.

[To be forwarded as soon as possible, without waiting for settlement of dispute.]

QUESTIONS.	ANSWERS.
1. Name of Trade affected	
2. Number of Firms involved [If an Employers' Association is concerned in the dispute, please give the name and address of its Secretary. If there is no such Association, please give the names and addresses of the principal firms involved in the dispute.]	
3. Cause or object of strike or lock-out (*Enclose copy of any application or Notice connected with the origin of the dispute.*)	
4. Date of the first day on which the workpeople were absent from work through strike or lock-out. (If notices were handed in, give also date of notice.)	

Occupations.	Men.	Women.	Apprentices or other Young Persons.
5. State occupations and number of workpeople (Unionists and Non-unionists) directly on strike or locked out.			
5a. State occupations and number of other workpeople (Unionists and Non-unionists) employed in above establishments who were thrown out of work owing to the strike or lock-out.			
Total Number of workpeople affected *			

* If any other workpeople were affected, respecting whom you can state no exact figures, please give, if possible, the name and address of some person who could do so:—

Information for the use of the Labour Department, Board of Trade, 44 Parliament St., S.W.

STRIKES AND LOCK-OUTS.

PART II.

[To be forwarded as soon as the dispute is terminated.] 1897.

QUESTIONS.	ANSWERS.
6. Date of termination of strike or lock-out, *i.e.*, the last week-day on which the workpeople were on strike or locked-out, or the date when the places of the strikers were filled up. (If there was no definite end to the dispute, please state approximately when it may be regarded as practically closed.	
7. Result of strike or lock-out (*Enclose copy of any printed or written agreement that may have been made.*)	
8. Describe the steps taken which resulted in the settlement, giving the names of any organizations or persons who assisted in bringing this about.	

9. If the result involved a **Change in the Rate of Wages or Hours of Labour**, give the following particulars for all workpeople whose wages or hours were changed, *whether Strikers or not*:—

Occupations affected by Changes in Wages or Hours.	Number of Workpeople whose Wages or Hours were changed.*	Date from which Change takes effect.	**Rate of Wages†** in a Full Week, exclusive of overtime.		**Hours of Labour** in a Full Week, exclusive of meal times and overtime.	
			Before Change.	After Change.	Before Change.	After Change.

* This is not necessarily the number on Strike or Locked out.

† When there has been a change in piece rates, please give the percentage increase or decrease in piece prices, and approximately the average earnings in a full week (exclusive of overtime) before and after change.

*Signature*_____

*Address*_____

*Date*_____

LABOUR DEPARTMENT, BOARD OF TRADE,
44 PARLIAMENT STREET, LONDON, S.W., 1897.

DEAR SIR, — The Labour Department of the Board of Trade is desirous of obtaining a complete and accurate record of Strikes and Lock-outs, and Changes in Rates of Wages and Hours of Labour in the United Kingdom as they occur, for publication in the Annual Reports presented to Parliament and also in the "Labour Gazette," which is issued monthly.

These statistics are collected and published by the Department in pursuance of the following Resolution adopted by the House of Commons on the 2nd March, 1886 :—

> "That in the opinion of this House immediate steps should be taken to ensure in this country the full and accurate collection and publication of Labour Statistics."

As the value of these statistics is greatly increased if the parties concerned co-operate with the Department by supplying accurate information, I should be glad if you would kindly answer as many as possible of the questions asked on the inner pages of this form so far as they relate to the _____

_____

If from any cause you are unable at present to answer the questions on Part II. of the form, will you be so good as to fill in and return Part I. at once, and send Part II. as soon as it is possible to do so.

I have to add that any information you may be good enough to furnish will be used solely for statistical purposes, and will not be published under your name.

A circular asking for similar particulars is addressed to the employer affected by this dispute.—Yours faithfully,

Chief Labour Correspondent.

The letter given with the later of these forms is a particularly careful one, showing the object of the inquiry, promising secrecy, and guaranteeing an impartial survey by the statement that similar forms are sent to employers and workmen. The forms addressed to employers are precisely similar in general appearance.

Change of form.

The main difference between the forms is that the later is divided into two parts, the first of which can be filled up directly work is stopped by a dispute, so as to give the Department a clue as to its magnitude and cause. The second part is detachable, and is to be preserved till the dispute is ended, and then forwarded. The advantages of this method are that the Department has early information as to the exact facts about the strike, and that the figures are given while the facts are fresh in the mind of the informant, whereas at the end of a long struggle, they might have been forgotten. Should the second part not be forwarded, the Department would of course write for it, or send a duplicate.

Question 1 on the earlier form is modified and split into two on the second. Question 2 on the later is simpler than the parenthesis of question 1 on the earlier, but asks for the more important information as to employers' associations, which will lead to the blank schedule being sent to the addresses given.

Question 2 of 1895, 3 of 1897, is the same on all four forms (the two to trade unions, and the companion forms to employers); it is not a statistical question, and probably leads to vagueness and to contradictory statements on the part of employers and employed; but the new parenthesis ("enclose copy, &c.") is important, for it leads to definite statements about which there can be no dispute.

The next question on the earlier form had of course to be altered for the new double sheet. Since the chief statistical information needed relates to the exact number of days' work lost, it is necessary to know exactly the date of the commencement of idleness; this day is therefore very carefully defined on the later form, as not that on which notices were sent in or any preliminary steps taken, but that of the actual commencement of hostilities. Question 6 (date of termination) is also carefully worded. The date of notices (question 4) gives useful subsidiary information.

There is considerable difference between question 5 and 5*a*

on the later and the corresponding question 7 on the earlier.

Number affected.

The only difficulty in using the information obtained from the new form arises at this point. The number affected by a strike, especially the number indirectly affected, changes continually, rising gradually to a maximum and then rapidly decreasing as the dispute draws to a close. The 1895 form did not give enough information, for the numbers at intermediate dates cannot be deduced from the numbers at the beginning and at the end, so that we have not the necessary data for determining what we chiefly want, the number of days' work lost (*i.e.*, the sum of the numbers of days lost by each person affected). In the case of a long dispute, however, this information is revised monthly at least, as is shown by the monthly report in the *Labour Gazette*.

The spreading of a strike.

The chief improvement in the new form consists in allotting separate spaces for different occupations. Several classes of workpeople will probably be affected in different ways by a strike in a complicated industry. Thus if the cotton spinners are on strike, very likely the carders will go out either on a grievance of their own or from sympathy. The spinners' assistants, the piecers, are at once thrown out of work, as are also the overlookers of the mules. As the strike continues all the departments of the spinning mill will be closed, one after the other. In the form four lines are allowed for those directly affected, eight for other classes unwillingly on strike. A great dispute, however, is not limited in its effect to the spinning mills. The supply of yarn falls off and the weavers are stopped; then the export trade is diminished, and dock labourers and sailors are thrown out of work, and so the influence of the strike spreads. It is out of the question to estimate completely these indirect effects; but in order to trace them as far as possible, space is given on the second form for the address of any one who can give information about them.

Question 6 on the earlier form, "Mode of settlement?" has been expanded considerably in the later one, since the question cannot well be answered in a single word, and the exact details are important for the non-statistical part of the inquiry. Question 5 in the earlier has also been altered; the important request for printed agreements is added, and the parenthetical part has been grouped with question 8 so as to form the new

question 9. This alteration, the same on employers' and workmen's forms, is worth special attention. The distinction between "directly" and "indirectly affected" is practically useless, and difficult to maintain in filling up the form. It is far more important to distinguish the different classes of workpeople, as can be done in the nine lines of the new form. Again, it is difficult to state the "total wages before and after," and the question leads to inaccuracies; the new question 9 is far better, for it is precisely that easiest to fill in, and most useful when done, and is in the exact form wanted by the Department for its register of changes of wages and hours. It is important in this question 9 to include all, whether on strike or not; hence we have the italics in the heading and a footnote to the second column. * This footnote could be improved, for at present the wording is a little obscure, and the notice might be put with advantage in the heading.

There remains a series of questions which have been dropped out in the later form. It may be supposed that it was found that the answers were not accurately given, that the inclusion of the questions overloaded the form, and by tending to inaccuracy in the answers led to inaccuracy in other details; while in cases when it was possible to obtain correct answers, it was found best to do so by other methods or a separate inquiry. There are two sets of questions: trade unionists are asked the amount they spent; employers the value of capital left idle.

Pecuniary loss.

Question 3 on the older trade-unionist form has nothing to do with the statistical inquiry. Question 9 simply affects the relation of unionists to non-unionists, may lead to exaggerated answers, is not wanted for tabulation, and is apart from the main inquiry. Question 10 belongs properly to a separate inquiry; the total might not be known to the secretary who fills in the form, and the amount expended by unions on "strike benefit" is compiled annually from other sources. The question is too complicated to be placed with advantage at the end of a long form. Question 11 is too vague to lead to the information wanted, though knowledge of the facts is needed. Question 12 is hardly likely to yield any results worth having, since all possible means have long been canvassed. The answers are, however, tabulated in the *Report on Strikes of* 1894 (C.—7901), which may be studied with advantage in connection with these forms, and some of

them may be quoted :—"Give in to the wants of the men, so that they are not extraordinary." "Abolish capitalism." "No means have yet been discovered." "Make all men Christian." "Fair argument." "A little more common honesty on the part of employers" (pp. 229-240).

The questions omitted in the later, but present in the earlier employers' forms are subject to similar criticisms. The subdivision of question 9, distinguishing summer and winter wages, and the separate columns for hours, are only in the later form. A comparison of these two forms with any number of the *Labour Gazette* and the *Annual Report on Strikes and Lock-outs of* 1894 will throw considerable light on the uses and difficulties of forms of inquiry.

SECTION 4.—STATISTICS OF ENGLAND'S FOREIGN TRADE.

The original schedules which lead to many other statistics are interesting, but limits of space must restrict us to one more typical inquiry, that which leads to our statistics of foreign trade.

In the population census the filling in of the form is compulsory and done by the householder; in the wage census the answers were voluntary and given once and for all by the employer; in the various inquiries undertaken by the Labour Department the answers are voluntary, but in many cases periodic, so as to become quasi-official. The method of collection of import and export statistics is a blend of all these.

The informants. There are three classes of persons who know the facts in question—the sender of the goods, the custom-house official through whose hands they pass, and the recipient or his agent. Circumstances decide that, in the case of exports from the United Kingdom, the exporter or his agent sends an account of the quantity and value of goods despatched to the Statistical Department of Customs; that, in the case of imports, the receiving-agent hands over an account of goods to be landed to the custom-house officials, who verify the account, roughly if the goods are duty free, carefully if they are liable to duty; and that, in the case of transhipment, the goods are treated in the same way as imports at the port of landing, and to some extent verified at the port of embarkation.

The blank forms, being filled in by officials as part of their duty, or by agents thoroughly used to the task, need no covering letter, and may be made as complicated as necessary; no questions are inserted but only blank tables. An examination of the forms in use will show what are included as exports and imports in the Board of Trade totals, and what is the total amount of information available for tabulation.

The quantities we wish to measure in this investigation are: the volume or weight and value of all goods which have an

The quæsita and data. exchange value, which leave our shores or reach them from without, subdivided as regards classes of commodities and countries of destination or origin; the values

being those at the times of loading or unloading. The quantities we can measure are sharply distinct from these, being the records of values and volumes which reach the Board of Trade. We should therefore examine the forms to decide—(1) What part of imports and exports are recorded; (2) whether the values are correctly given, (3) the quantities accurately registered, (4) the commodities accurately defined, (5) the countries of origin and destination accurately distinguished in the returns.

On reaching port the ship's master has to send in an account, of which the following is an abridged specimen :—

Examples of information.

If Sailing Vessel or Steamer — STEAMER. — Official No. / No. of Register, / Date of Registry,

No. 1. REPORT No. 980.*

Port of X.

Ship's Name.	Tonnage.	British or Foreign. If British, Port of Registry; if Foreign, Country to which she belongs.	Number of Crew. British Seamen.	Number of Crew. Foreign Seamen.	Name of Master, and whether a British or Foreign Subject.	Port or Place from whence arrived.
Marianne.	700	BRITISH.	12	—	H. Hind.	Havre, France.
		Total..				

CARGO.

1. Name or Names of Places where laden in order of time.	2. Marks	3. Nos.	4. Packages and Description of Goods, Particulars of Goods stowed loose, and General Denomination of Contents of each Package of Tobacco, Cigars, or Snuff intended to be imported at this Port.	5. Particulars of Packages and Goods (if any) for any other Port in the United Kingdom.	6. Goods (if any) to be Transhipped or to remain on Board for Exportation.	7. Name of Consignee.
Havre, France.	Paris	to	London.—600 pkgs.	Fruit and Perishables.		Smith.
			68 pkgs. Merchandise.			
	COK	1392				
	AE	495/6				
	KG	340/9				
	FOT	1/50				
	AJ	3/6				
	CK	1				
	AC	10				
	KL	40	70 cases Wine.			,,
If any wreck fallen in with or picked up, to be stated.	ACD	20				
	WD	166	5 cases Woollens in transit to Liverpool.			,,
	O&D	1	1 case Brandy.			,,

STORES.

Surplus Stores remaining on board, viz. { 3 lb. Cigars. / 4 lb. Tobacco.

Number of Alien Passengers (if any) - Nil.

Pilot's Names - - - - - ———

At what Station Ship lying - - - South Quay.

Agent's Name and Address - - - C. J. C.

I declare that the above is a just report of my Ship and of her Lading, and that the Particulars therein inserted are true to the best of my knowledge, and that I have not broken Bulk or delivered any Goods out of my said ship since her departure from Havre, the last Foreign Place of Loading.

(Signed) H. HIND, *Master.*

Signed and declared this 13th day of October 1896

In presence of

(Countersigned)

Collector.

* *i.e.*, 980th ship at X. since 1st January.

The goods for quick transit are passed at once, and a special form is sent to the Customs Establishment similar in character to that on p. 67. The remaining goods are treated either as dutiable or as duty-free articles. In the list before us, ten cases of wine are entered for home use, and an account is sent in to the Statistical Office; sixty cases are warehoused and another account (as to quality, quantity, and value) is sent in; the whole are registered as imports. Twenty of the warehoused cases are removed to another port and re-exported; an account is sent, and they are entered as exports of foreign goods. Twenty are put on board ship as stores at the original port, and twenty more removed to another port for the same purpose, and of this the central office takes no account; the remainder are removed to another warehouse, still in bond, and on leaving that will be treated in one of the four ways just mentioned. Other dutiable articles are treated in the same way.

Dutiable goods.

Goods not sufficiently described or not answering to their description are opened, their contents entered on a "bill of sight," and an account sent in. Private effects are separately examined, being described on a "sufferance" form; if they are *bona-fide* personal goods no record is kept of them, except in the case of dutiable goods, which are treated as ordinary imports. If the dutiable goods are concealed, either among private effects or merchandise, and forfeited, they are not reckoned as imports.

Examination of goods.

Bullion is entered on a separate form and kept distinct throughout the accounts.

The duty-free goods, if for transhipment at another port, are sent there under seal, and barely examined; they are treated at the central office in the same way as dutiable transfer goods. The remaining free goods, which in general form the bulk of the cargo, are entered on such a form as follows, which is worth notice, for it is a specimen of the rough material from which our foreign trade figures are evaluated.

Free goods.

ENTRY FOR FREE GOODS.

This space is for the use of the Officers of Customs.

Port_____

Dock or Station_____

Importer's Name_____ (No._____)

Examination.	Ship's Name.	Master's Name	Rotation No.	Date of Report	Port or Place whence
	Marianne.	H. Hind.	980.	13/10/96.	Havre, France.

Marks and Nos.	No. of Packages and Description of Goods, in accordance with the Official Import List.	Quantity.	Value, £.
COK 1392	One Goods Manuf. N.O.E. Billiard Cue Tips	...	28
AF 495/6	Two Leather Shoes	10 doz. prs.	58
KG 340/9	Ten Cotton Manuf. Trimmings - Embroideries	...	140
		...	280
	Piece Goods, not Muslins	300 yds.	8
FOT 1/10	Ten Gloves of Leather	11,240 doz. pr.	12,316
„ 11/5	Five Silk Broad Stuffs	...	10,400
„ 16/20	Five Works of Art—		
	Plaster Casts	...	380
	Statuary	...	1,280
	Pictures by Hand	3	10,200
„ 21/5	Five Books Bound	4 cwt.	300
„ 26/30	Five Bronze Manuf. Ornaments	3 cwt.	38
„ 31/5	Five Metal Manuf. Ornamental Brass-headed Nails	4 cwt.	24
„ 36/40	Five Silk Manuf. Dresses, Mantles, Trimmings	...	1,816
„ 41/50	Ten Goods Manuf. N.O.E.—		
	Fancy Goods	...	110
	Horseless Carriage	...	160
	Brushes	...	78
	Glue	...	110
	Billiard Chalk	...	12
	Hardware	...	116
AJ 3/6	Four Stationery Ink	...	48
CK 1	One Iron and Steel Manuf. Machinery, British, returned	3 cwt.	24

I enter the above goods as free of duty, and declare the above particulars to be true.

Dated this 13th day of October 1896.

(Signed) J. JONES,
Importer or his Agent.

Verification of data.

The information so received is usually accepted at the central office without inquiry. It frequently happens, however, that the form is not properly filled in by the agent, the values often being omitted. When this is so, it is the duty of the clerk at the port of entry to fill in the value, in accordance with a list of current prices with which he is

provided. It may happen that he has to appraise the goods on inspection, a process leading in some cases to great error, which is enhanced when not even the quantities are given. When there is a palpable error or omission in the form, or when the price appears out of the common, a query is sent from the central office to the port: *e.g.*, with reference to such a form as that just given, the following correspondence might arise:—

1. Pictures by hand, £10,200. Explain high value. *Answer.*—Correct; invoice was seen; pictures by Millet.

2. Books bound: is weight or value incorrect? *Answer.*—Both correct; advice seen; old and valuable books.

3. Goods entered as "goods manufactured, chip plaiting": explain nature, and state if description is correct. *Answer.*—Correct; wood shaving plaited and occasionally mingled with horse-hair, &c.

4. Potatoes, 40 cwt., £62. Weight or value? *Answer.*—Value correct. Weight should be 400 cwt.

Thus any unusual entries are liable to be checked and verified.

In the case of goods not easily valued, or of miscellaneous goods not easily tabulated, errors must arise in this way; and another error may enter if a clerk, who does not wish to receive too many queries from head-quarters, enters at ordinary rates goods of exceptional value; but when staple commodities and large quantities are involved, all the persons concerned will be familiar with the forms they have to fill, the prices will be known, and so in important cases errors will be at a minimum. The import total values, therefore, are the sum of many quantities of various degrees of accuracy, and it is not difficult when looking through the list of items in the annual report to see which are specially liable to error. Such commodities as old books, works of art, goods where sale depends on the fluctuations of fashion, racehorses, and so on, have values varying from day to day, and their exact value in the balance of imports and exports cannot be determined.

Possibility of errors.

The quantities and values of exported goods are filled in by the shipper or agent, and sent to the central office within six days of the ship's clearing. The following is an abridgment of the form used:—

Exports.

*SPECIFICATION FOR BRITISH AND IRISH GOODS ONLY.

Port of X. Ship Name—"Marianne." H. Hind, Master, for Havre.

* The Specification of Goods exported must be delivered to the proper Officers of Customs within six days from the time of the final clearance of the ship, as required by the Customs Laws.

Marks.	Numbers.	No. and Description of Packages.	Quantity and Description of British and Irish Goods, in accordance with the requirements of the Official Export List.	Value.
KCL	641/2	Two bales	Woollen Heavy Cloths. 1,400 yards - - - -	£340
CKD	140/1	Two cases	Steel Manuf. Blades. 3 cwt. - - - - -	24
RMO	10/12	Three crates	Steel and Iron Manuf. Threshing Machine. One - -	380
CL	140	One case	Cotton Piece Goods, Printed and Dressed. 1,800 yards -	50
			Total - -	£794

I declare that the particulars set forth above are correctly stated.

(Signed)_____ *Exporter or Agent.*

Address—Harbour Buildings.

Dated 13th October 1896.

(Countersigned)_____

Officer of Customs.

The forms for British and Irish goods are distinct from those for foreign, free and duty-paid, goods; and there are distinct export forms for transhipments, which have already been registered as imports. In these cases the specification and quantities are likely to be correct, but there are causes which may falsify the values. If they are to be subject to an *ad valorem* duty, they may be undervalued; if they are adulterated goods, masquerading as genuine, they may be over-valued. It seems hardly possible to estimate these errors.

Definition of official imports and exports.

We are now in a position to define imports and exports according to their meaning in the Board of Trade Returns; as, for instance, when for 1895 the value of imports is stated as £416,000,000, and of exports as £285,000,000, of which £60,000,000 are re-exports of foreign or colonial goods.

This total for imports includes all goods landed through the custom-houses, including goods immediately shipped as stores, or returned from customers unused. Goods immediately re-shipped at the same or another port, or held in bond and then re-shipped, are included both as imports and exports. Bullion is not included, being given separately, nor cargo unlanded and so reported, nor personal luggage or private effects, except when duty is charged. The value reckoned is the nominal exchange value when or just before they are landed; that is, their value is already increased by freight, but not increased by duty.

The total for exports includes all goods entered on ships' bills of lading, does not include ships' stores or passengers' luggage, nor cargo unlanded and so reported, nor bullion, which is given separately. The value is reckoned at the time they are put on board. Ships leaving our shores to be sold to foreigners are now included.

The treatment of coal throws light on this paragraph. Coal taken for use on the voyage is registered, but not included among exports; coal as cargo is included.

Among exports not registered are cash taken privately and personal effects; among imports not registered are smuggled goods, and cash and personal effects.

For the causes and extent of the resulting differences between imports and exports, Sir R. Giffen's two papers * on the subject should be consulted.

* *Essays in Finance*, Second Series; and *Journal of the Royal Statistical Society*, 1899.

CHAPTER IV.

TABULATION.

CHAPTER IV.

TABULATION.

LEAVING now the consideration of blank forms of inquiry, let us turn to the methods by which our data, accumulated on these forms, can be tabulated. At first sight the tabulation of so many million census forms, so many schedules of wages, and so many lists of goods imported, seems mere office work, to be done mechanically,* only requiring accuracy and not subject to scientific analysis. Tabulation does, indeed, involve a great deal of automatic labour; but the determination of the exact form of the table and the choice of the headings to which the totals shall correspond task the administrative statistician, and are worth the closest study.

The function of tabulation in the general scheme of a statistical investigation is sufficiently definite; it is to arrange in easily accessible form the answers to those questions with which the investigation is concerned. If it is required to know, for instance, the number of persons of each sex and age-group in all the districts of the country, the figures in the table must show these numbers. Or, to take a less definite problem, we want all the information possible as to labour disputes. In studying the forms issued by the Labour Department, we have seen that the information which can be obtained is not precisely that which we require. The problem then is so to tabulate our information that our totals may give answers as near to our requirements as possible, and it can easily be found by experiment that the way to do this is by no means obvious.

The function of tabulation.

Not only must the figures be grouped so as to answer the questions put forward in the original scheme, but if the information is of wide and varied interest, as in all the inves-

* An account of Mr Hollerith's electrical tabulating machine, used in the XIth Census of the U.S.A., will be found in Dr Bertillon's *Cours Elémentaire*, p. 579 *seq.*

tigations so far considered, the data must be studied from many points of view, and tabulated so that students in all branches of knowledge may be able to extract from our tables the information they require. Thus the population census is used by the financier, the legislator, the merchant, and the commercial traveller; political economists turn to it for light on the development of industry, and on the change of numbers in each trade; those interested in social questions will study the ages and sex-distribution in various districts or occupations; the sociologist and biologist will need accurate information as to the growth of population and the change of age distribution.

To take more specific points, the blue-book which contains the tabulation of foreign trade statistics will be expected to show how our trade with each country is developing, whether we are holding or improving our position in certain markets; whether we are exhausting our supply of raw materials; whether some new commodity is yet of importance. It must be remembered that the original material is not accessible to the public, that they are dependent on the information extracted for them, and that, though it would be possible to turn through all the forms for special data, yet the labour needed would be prohibitive, while a little more detail in the tabulation might easily have isolated the information needed.

For convenience, the methods of tabulation may be divided into three groups: A. The simple statement of totals of persons or things which satisfy given conditions, such as the number living in a town, or the total value of imports from France; B. The grouping of a great number of units in relation to some particular property possessed by all, with the object, not of answering assigned questions, but of putting the material in a form ready for use in further investigations—*e.g.*, the population according to ages, or wage-earners according to the value of their wages; C. The tabulation of non-numerical answers in suitable groups to give a view of the whole—*e.g.*, the causes of strikes or the state of employment. The division between groups A and B is not always definite.

Three groups of tabulations.

In the tabulation the convenience of the reader must be studied. The table must be so arranged that any totals required can instantly be found. This is to a great extent a question of

typography, the use of suitable founts for figures and headings, and also of the choice of the right shape and size of page. Supposing the best possible choice made in these respects, our rule will then be to get the maximum amount of information into the minimum space.

Classes of tabulation.

GROUP A.—Thus we can have SINGLE tabulation, answering one or more groups of independent questions, as :—

NUMBER AND MEMBERSHIP OF TRADE UNIONS.*

Year.	Number of Trade Unions at end of Year.	Total Membership of these Unions at end of Year.
1896	1,317	1,493,375
1897	1,307	1,611,384
1898	1,267	1,644,591

DOUBLE tabulation shows the subdivision of a total according to two categories, in the following example according to sex and age :—

CLASSIFICATION OF PAUPERS IN IRELAND.—Total Numbers who received Relief during the Year ended Lady Day 1892.†

Ages of Persons Relieved.	Males.	Females.	Total.
Under 16 years - - -	44,391	43,648	88,039
Of 16 and under 65 years -	132,370	79,045	211,415
Of 65 years and upwards -	35,121	45,668	80,789
All ages - - -	211,882	168,361	380,243

* Compiled from the *Sixth Annual Abstract of Labour Statistics*, p. 1.
† *Ibid.*, p. 102.

More information may be included thus :—

CLASSIFICATION OF PAUPERS IN ENGLAND AND WALES. — Total Numbers who received Relief during the Year ended Lady Day 1892.*

Ages of Persons Relieved.	Indoor.	Outdoor.	Total.	Metropolis.	Other Parts of England and Wales.
Under 16 years	141,782	441,805	553,587	100,671	452,916
Of 16 and under 65 years	232,284	385,299	617,583.	148,066	469,517
Of 65 years and upwards	114,144	287,760	410,904	64,779	337,125
All ages	458,210	1,114,864	1,573,074	313,516	1,259,558

A TREBLE tabulation can be used, subdividing the total into three distinct categories, with cross totals for each group. Thus the following table gives separate divisions according to age, sex, and district; percentage lines, in a distinct type, are also introduced :—

* *Ibid.*, p. 101.

Classification of Paupers by Age, Sex, and Locality.*

Total Numbers who received Relief in England and Wales during the Year ended Lady Day 1892.

Ages of Persons Relieved.	Metropolis.				Other Parts of England and Wales.				Totals. England and Wales.			
	Males.	Females.	Total.	Percentage at each age.	Males.	Females.	Total.	Percentage at each age.	Males.	Females.	Total.	Percentage at each age.
Under 16 years† - -	...	...	100,671	*32.1*	...	...	452,916	*35.9*	...	...	553,587	*35.2*
Of 16 and under 65 years	74,207	73,859	148,066	*47.2*	202,180	267,337	469,517	*37.3*	276,387	341,196	617,583	*39.3*
Of 65 years and upwards—												
Number - -	27,238	37,541	64,779	*20.7*	136,392	200,733	337,125	*26.8*	163,630	238,274	401,904	*25.5*
Percentage of sexes	*42.0*	*58.0*	*100*	...	*40.4*	*59.6*	*100*	...	*40.7*	*59.3*	*100*	...
All Ages - - -	...	...	313,516	*100*	...	...	1,259,558	*100*	...	...	1,573,074	*100*

* *Ibid.*, 101.

† The returns do not distinguish sex under 16 years.

The same process can be further extended: the example in the table opposite shows an arrangement for a QUADRUPLE tabulation, distribution by district, date, sex, and occupation, with subsidiary information; but it is generally better to use two or more tables than to increase the complication, unless it is necessary to bring several categories into close relation. Suitable varieties of type will often make comparisons easy in a very complex table.

Looking now at the census householders' schedule (p. 23), it will be seen that there are about twelve different items of information about each person: county, town, parish, position in family, civil condition, sex, age, occupation, industrial position, infirmity, birthplace, and house-room. These could be tabulated in 66 different single, 220 double, or 495 treble tabulations, so that there is plenty of scope for choice.

Tabulation of census material.

To fix our ideas, we will take occupation as the main sub-division, and examine Mr Booth's use of the census returns, say for London Printers.*

Mr Booth's tabulation.

First he gives a treble classification—occupation, sex, and age—using columns 4, 5, and 6 of the schedule.

Census Divisions, 1891.	Females.	Males.			Total.
	All Ages.	-19.	20-54.	55-	
1. Printer - - -	1,316	9,988	21,784	1,921	35,009
2. Lithographer, &c. -	809	757	3,037	437	5,040
Total - -	2,125	10,745	24,821	2,358	40,049

Then follows a single table, district and numbers, using the information on the back of the schedule.

Distribution.

E.	N.	W. & C.	S.	Total.
5,884	9,835	7,577	16,753	40,049

* *Life and Labour of the People*, vol. vi., p. 189.

NUMBER OF MALES AND FEMALES of all ages engaged in certain occupations in ENGLAND AND WALES and SCOTLAND in 1881 and 1891.

	ENGLAND AND WALES.								SCOTLAND.								TOTALS. 000's omitted.					
	MEN.				WOMEN.				MEN.				WOMEN.				1881.			1891.		
	1881.		1891.		1881.		1891.		1881.		1891.		1881.		1891.							
	Numbers employed. 000's omitted.	Per 10,000 Males above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Males above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Females above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Females above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Males above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Males above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Females above 10 years of age.	Numbers employed. 000's omitted.	Per 10,000 Females above 10 years of age.	MEN.	WOMEN.	ALL.	MEN.	WOMEN.	ALL.
Cotton Manufactures -	185	199	212	202	303	303	332	290	3	23	6	42	26	173	13	82	188	329	517	218	345	563
Woollen Manufactures	93	100	102	96	123	123	132	115	14	104	14	94	16	111	18	115	107	139	246	116	150	266
Flax and Linen Manufactures - - -	4	4	3	3	8	8	6	5	8	67	7	48	20	136	19	120	12	28	40	10	25	35
Boot and Shoe Manufactures - - -	188	202	202	191	36	36	46	40	22	170	19	129	2	14	3	19	210	38	248	221	49	270
All occupations - -	7,753	8,324	8,866	8,374	3,402	3,405	3,945	3,442	1,092	8,302	1,204	8,525	484	3,163	544	3,400	8,845	3,886	12,731	10,009	4,489	14,498
Population above 10 years of age - -	9,314		10,592		9,992		11,461		1,313		1,446		1,462		1,599		10,627	11,454	22,081	12,038	13,060	25,098

To face page 78.

Three simple tables are then given, relating to heads of families, using columns 2 and 4 (sex), 2 and 10 (birthplace), and 2, and 7, 8, 9 (industrial status).

His next table uses columns 2 and 6, and is as follows :—

	Total Population Concerned.				
	Heads of Families.	Others Occupied.	Unoccupied.	Servants.	Total.
Total - - -	18,048	16,060	47,257	854	82,219
Average in Family -	1	.89	2.62	.05	4.56

The next table (not here given) is a single classification according to number of rooms and servants, a most ingenious indirect use of the scheduled information; and the last is an example of the legitimate use of a quadruple tabulation—occupation, industrial status, sex, and age—given on the next page.

STATUS AS TO EMPLOYMENT (according to Census Enumeration).

CENSUS DIVISIONS (1891).	EMPLOYERS.		EMPLOYED.			Neither Employer nor Employed.		Total.
			Males.		Females of all Ages.			
	Males.	Females.	Under 20.	Over 20.		Males.	Females	
1. Printer	827	39	9,988	22,565	1,266	313	11	35,009
2. Litho., Copper, and Steel-plate Printer	177	2	506	2,571	88	153	6	3,503
2. Map and Print Colourer and Seller	17	...	49	175	72	62	12	387
2. Ticket and Label Writers	36	3	202	169	619	114	7	1,150
Total	1,057	44	10,745	25,480	2,045	642	36	40,049
	1,101		38,270			678		
			Proportion of Employers to Employed--			1 to 35.		

It would be difficult to find a better example of tabulation of a great multitude of details to serve a special purpose. The census authorities had in many cases not tabulated the necessary details, and it was necessary to turn through the original schedules to get at the facts. For such work as this, the function of tabulation is simply to provide the answers to definite questions. Thus the census reports show how many persons of each sex and age-group belong to certain industries in certain places, in a quadruple tabulation extending over many pages, each page relating to one district, and this table may be used for accomplishing many separate purposes: each item is already a total ready for use. It is impracticable from limits of time and space, even if it were desirable, to tabulate all the possible groups of qualities which can be made from the twelve statements on each census form; a good tabulation will aim at providing only those statements which are of practical use. Thus many simply descriptive totals are given, such as the numbers of each sex and age in each parish in the United Kingdom, to serve primarily for administrative purposes; and many statements which will afford the economist and sociologist the opportunity of tracing the progress of industries, of studying the ages of workpeople in different occupations, the changes in age-grouping of the nation; and some further tables might be given to throw light on problems of cause and effect, such as the average ages in town and country, the connection between infirmities and occupation, or the ages of marriage in various districts or industries.

The census tabulations.

It is interesting to open one of these great tables of figures, such as are generally to be found forming the bulk of a blue-book, and taking a figure at random, ask "Why is this figure printed, what question does it answer, to whom can it give information?" For instance, in the *Eighth Report on Trade Unions*, p. 257, we find that the United Brickworkers' and Brick Wharf Labourers' Union spent £20 on funeral expenses in 1894, an average of 3s. 7¼d. per member. As an isolated statement this may interest a very small number of persons; but that small number has a right to expect that they shall find the figures relating to their union tabulated in a general official book; to them it may be as important as the item, on the same page, of £5,481 spent by the Boilermakers. From this point of view, the question of inclusion of

Minutiæ.

such small items is simply one of space. If space is limited, a selection would be made of larger quantities only, as being likely to concern more people.

Importance of raw material.

But there is a reason of quite another character for printing such items as these. The raw material, on which the totals in such tables are based, is not accessible to the student except by means of this Report. Now, the compiler of these statistics cannot know from what particular point of view they will be studied. It may be desired to examine and group trade unions according to their expenditure on different items, to study their history, classifying them as fighting organisms and as friendly societies. The tabulations needed cannot well be foretold. The material is therefore given in the rough, in order that the tabulation may be made by each student according to his needs. At the same time the most suggestive totals are given as one of these possible methods of tabulation; and in the summary of such a report, the items are retabulated, the rough material being omitted, in those ways which the editor thinks most useful.

Selection of raw material.

When space is much too limited for any publication *in extenso* of the items, a careful selection must be made of those to be printed; and it is this selection that is generally open to most criticism. Owing to the great admiration for uniformity generally to be found in the official mind, valuable space is wasted on such statements as *—

COVENTRY: 1891	MALES.
SHIPWRIGHT: Ship, Barge, &c., Builder (Wood)	1
,, ,, ,, ,, (Iron)	0

while all the males—masters, traders, skilled workmen, labourers, errand-boys—engaged in the cycle trade in Coventry are included in—

Bicycle, Tricycle—Maker, Dealer	3,854

In such cases, two useful rules might be applied: omit all numbers under, say, 500 when by so doing a *line* of print would be saved; and give all numbers over 10,000 correctly only to the nearest 100, and so for other digits in proportion, thereby

* Census Report.

reducing the width of *columns* of print. If, for example, we knew to the nearest 100 the exact numbers in each district and occupation in which as many as 1,000 were employed, our knowledge would be as complete as we needed; and it is doubtful whether the space occupied by this tabulation would be more than that already dovoted to the subject. In many cases, on the other hand, it is essential to have the raw material quite unchanged. Each tabulation must be judged on its own merits.

Economy of space.

It may be useful to take a particular group of answers, and discuss what tabulations will throw most light on the questions at issue. The Poor Law Commissioners of 1833 collected information from a thousand villages in England and Wales on the following six points among others: the wages of an agricultural labourer in summer and in winter, both with and without the inclusion of beer as part payment, his annual earnings, and the subsidiary earnings of his wife and children. It may be supposed that the chief object of the Commissioners was to find whether the labourers' families earned enough for their support, and what proportion was earned by the wives and children.

Tabulation of the Poor Law Returns, 1833.

The following scheme of tabulation would show in what *counties* the labourer was badly off:—

County.	Average Annual Earnings of		
	Man.	Family.	Together.

The counties might be taken in alphabetical order for convenience of reference, or in geographical order with subordinate averages for groups (*e.g.*, Eastern: Norfolk, Suffolk, Essex); or the counties might be arranged in the order of the total earnings, so that it could be seen at a glance in which counties the labourers were worst off.

To show the number of villages, county by county, in which the earnings were below a certain minimum, or within certain limits, the following table might be used:—

ANNUAL EARNINGS OF MEN AND FAMILIES.

	Number of Villages in which the Total Earnings averaged							Average Earnings in County of		
	£25.	Above £25 and not above £30.	Above £30 and not above £35.	Above £35 and not above £40.	Above £40 and not above £45.	Above £45 and not above £50.	Above £50.	Man.	Family.	Together.
IN NORFOLK -	0	1	3	6	4	3	2	£30	£11	£41
Percentages of Total Number of Villages -	0	5	16	31½	21	16	10½	73	27	...
IN SUFFOLK -	0	3	4	5	3	2	2	£28	£11	£39
Percentages of Total Number of Villages -	0	16	21	26	16	10½	10½	72	28	...
IN ESSEX -	1	3	6	7	10	3	1	£28	£10	£38
Percentages of Total Number of Villages -	3	10	19	23	32	10	3	74	26	...
In Eastern Counties	1	7	13	18	17	8	5	£28 10	£10 10	£39
Percentages of Total Number of Villages -	1	10	19	26	25	12	7	73	27	...

This table can be used in the above complex form or simplified. The number of subdivisions of money to be distinguished depends on the space at disposal and on the number of villages which would be entered in each. A table in which most of the entries are 1 or 0 is open to criticism. In the above table the villages are too few to allow accuracy in percentage.

It will be seen that this table would furnish the answer to almost all questions which could be put as to total earnings.

Tabulation to show correlation. For instance, if we wish to see the relation between total earnings and the family's subsidiary contribution, we should look at the smallest totals in the last column and see if they corresponded with the largest percentage of family earnings. If we found signs of correspondence we should re-arrange the counties in the order of these subsidiary percentages, and see if they were approximately in order of total earnings also. This is an example of tabulation to show correlation, the correspondence in the occurrence of two sets of phenomena.

Another important group of questions arising in connection

with these tables is: What is the relation between weekly wages and annual earnings, and what proportion of the wage is generally paid in kind? We shall not now require the statements as to subsidiary family earnings. In records of agricultural wages the most common statement is, *e.g.*, "wages in this district are from 10s. to 12s. a week." Now, a farm labourer does not generally earn as much in winter as in summer, because wages are reduced to correspond to the smaller amount of work necessitated by failing light; from this cause annual earnings will be less than the weekly wage multiplied by 52. Besides this wage he generally receives special money at hay and wheat harvests, and also many payments in kind, such as daily beer, house and ground at reduced rent, and other privileges. It is generally best to value all these, and compute his earnings thus:—

Wages and earnings.

	£	s.	d.
10s. for 38 weeks -	19	0	0
12s. for weeks 9 (summer)	5	8	0
Hay harvest, 1 week -	0	15	0
Wheat harvest, 4 weeks -	5	0	0
Beer, 1s. per week - -	2	12	0
Cottage and ground -	5	0	0
Other perquisites - -	1	5	0
	£39	0	0 = 15s. per week.

In this case earnings are 50 per cent. above the general weekly wage. An estimate of this nature has been made by the late Mr Little for each county for 1867-70 and 1892. We can tabulate the figures for 1833 in the same way for comparison, in geographical order and with the county as unit. We must first consider the question, Has beer been at all generally replaced by money? We can tabulate the figures as follows to answer this:—

Beer.

1833.						1892.	
1. County.	2. Average Summer and Winter Weekly Wages.	3. Average Earnings per Week.	4. Difference	5. Number of Villages where Beer is given.	6. Proportion to Total.	7. Difference between Wage and Earnings.	8. Excess of 4 over 7.

In column 2 should be given the county average of the wages stated, without making any cash allowance when beer is given. Then if money has been replacing beer, we should find that in those counties where beer was most often given, wages had risen relatively to earnings more rapidly than in the counties where free beer was rare. Columns 4 and 7 show the differences for the two dates. When the entry in column 4 is greater than the corresponding entry in column 7, kind has been replaced by money. These excesses would be given in column 8. If money has replaced beer, the counties which have the greatest entries in column 6 should also figure high in column 8, and *vice versa.*

Winter and summer wages.

The question, Are winter wages generally below summer wages, and by how much? can be answered by the following scheme of tabulation, which uses the data not employed in the previous tables:—

Counties.	Average Weekly Wage in		Number of Villages where the Excess of Summer Wages over Winter was					
	Summer.	Winter.	Nothing.	6d.	1s.	1s. 6d.	2s.	More than 2s.
	s. d.	s. d.						
Norfolk - - -	11 2	10 3	13	2	3	2	5	3
Percentage of Number of Villages included - - -			*46*	*7*	*11*	*7*	*18*	*11*
Suffolk - - -	10 2	9 8	24	0	6	1	2	1
Percentage of Number of Villages included - - -			*70*	*0*	*18*	*3*	*6*	*3*
Essex - - -	10 9	9 10	22	0	11	0	5	4
Percentage of Number of Villages included - - -			*52*	*0*	*26*	*0*	*12*	*10*
Eastern Counties	10 6	9 11	59	2	20	3	12	8
Percentage of Number of Villages included - - -			*57*	*2*	*19*	*3*	*12*	*8*

These examples do not quite exhaust the useful tabulations of these groups of figures, for we have not yet examined the distribution of wages, that is the relative numbers paid at

different rates. These returns do not, however, illustrate such a tabulation well, for we are not told the rates paid to individuals, but only the rate prevalent in the villages.

GROUP B.—The grouping according to wages affords an example of the second method of tabulation. We have now no definite questions to answer, as in the method so far discussed, but a more general problem: given a mass of data, it is required to tabulate it, so as to present the maximum amount of useful information. Our raw material is so many thousand isolated statements, which must be focussed, made to present definite meaning, and worked up so as to be useful for future comparison.

Statistics whose purpose is not definite.

Some investigations are undertaken not to answer any definite questions or to throw light on any given problem, but to collect information which, though it has no immediate use, is likely to be needed ultimately by many investigators occupied with various questions. Such is a wage census. So long as we have no sufficient account of wages, we are badly informed as to one of the most important measurements of the social body, and economists and statisticians are continually hindered by the want of data essential for their work; but the census has no immediate practical use, for knowing the height of wages does not help us directly to regulate that height. In such an investigation our object will be to examine the figures, and give all the groupings and averages which seem likely to be useful for any purpose; and while doing this we shall imperceptibly pass to a different class of investigation; we shall be finding a structure underlying our multifarious details; we shall find that the chaos, which our figures present at first sight, obeys laws; we shall be making a visible outline, and giving a definite shape to our apparently featureless mass.

The complete discussion of this problem belongs to a later chapter; but the tabulation can be begun without special technique. The examples taken will relate chiefly to wages, but the methods are quite general.

In the American *Report on Wholesale Prices, Wages and Transportation* of 1891, the wages of some 10,000 persons are detailed. It is proposed to consider their tabulation as a homogeneous group. The results are given on pp. 91-2. In the original publication the wages are given to half a cent; in the second column, on p. 91, the numbers of

Selection of limits of groups.

wage-earners are given in 10-cent groups, from $.25 to $.34, $.35 to $.44, and so on, those earning wages exactly at the dividing points being always placed in the division below. Notice that the average wage of such a group as $2.15 to $2.24 is not $2.20 if the wage-earners are evenly distributed cent by cent, but the average of $2.15, $2.16, . . . $2.24, *i.e.*, $2.195.

Looking at column 2, it will be seen that the figures present no order, follow no rule; no structure has yet been found, our divisions are too narrow for our material.

Now group the wage-earners with wider limits, as in column 6, where the numbers earning in half-dollar groups are given; we have here a nearly regular sequence of numbers falling after the maximum in the second group. Going back to narrower limits, to find exactly at what divisions this regularity is first in evidence, we have in column 4 the numbers in 20-cent groups which show considerable, but not absolute regularity. The numbers in 30-cent groups * are successively 75, 355, 674, 1,242, 740, 660, 343, 310, 180, 181, 233, 32, 82, 3, 4, 8, 1, almost completely regular except for the large group at $3.50.

The question as to which of these groupings should be selected is to be decided by the number of separate items the eye can instantaneously grasp. In looking at the 25 numbers in the 20-cent groups, or the 18 in the 30-cent, the meaning is lost in a maze of figures (though as many details as these could be properly shown in a diagram), but the 11 numbers in the half-dollar groups are easily comprehended.

Stated in words, the result of our tabulation (column 7) is that 6 per cent. of the wage-earners made from $.25 to $.74, 29 per cent. from $.75 to $1.24, and so on.

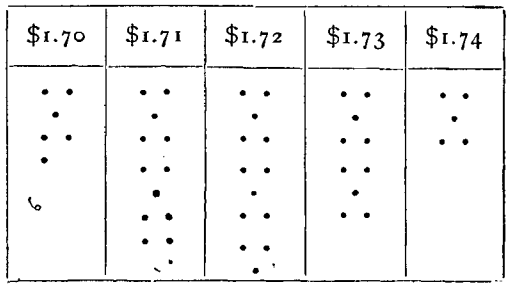
For the practical work of the tabulation from the original figures, we should take ruled sheets, enter at the head of successive columns certain wage limits, and turning through the items enter each wage by a dot in its appropriate column, grouping them in fives and tens, to facilitate addition.

Practical tabulation.

From the preceding paragraphs it is clear that we do not need to take separate columns for each cent from $.25 to $5.35 for tabulation, but a little consideration is necessary to see how minute the limits should be to give the correct average.

* *Vide* p. 121 *infra*.

Suppose the entries in cent groups to be:—



The average of the wages so entered can be quickly calculated as $1.718.

If, on the other hand, we put all the 46 entries as simply "between $1.70 and $1.74," or more exactly "as much as $1.70 but less than $1.75," we should naturally take them to be all (for purposes of averaging) at the middle point of this group, viz., $1.72.

If we have a sufficient number of items, the differences between the average assumed and that calculated for each group will be very slight. This is seen on p. 91; column 8 gives the averages calculated from the entries in 10-cent groups, while column 9 gives them on the hypothesis that for purposes of averaging the numbers in the half-dollar groups may all be taken at the middle points of their groups. The difference is greatest in the first and last, the smallest groups. The general average obtained from column 9 is $1.70, which is the nearest round number to the true average $1.73. Hence, for the purpose of obtaining the general grouping and average, we need only take 11 half-dollar columns for marking in our items.

For other purposes it may be advisable to work more minutely; for in the lowest group, we shall wish to know how many are earning $.25, $.30, $.35 separately, for 5 cents is a perceptible difference on 25 cents. At the top also it may be useful to know the exact wages.

More minute entries again will be needed for the second method of tabulation, which is as follows:—Suppose all the wage-earners to be arranged in order of the magnitude of their wages, those at $.25 at one end, those at $5.75 at the other. Note the wages of men at given points in the row. The lowest wage is $.25; one-tenth of the way along,

The Galtonic method.

that of the 512th worker is between $.85 and $.95, . . . : half-way up the wage is $1.50. The figures at each tenth are given on p. 92. By this means we get a very vivid idea of the distribution according to wages.

These numbers cannot be obtained accurately if we have only entered the details correct to half-dollars, but can be found from the 10-cent grouping, which is therefore the classification to be adopted. We must first determine in which of the small groups the men one-tenth, two-tenths . . . up the group lie, and then estimate their position inside the smaller group. Thus, if we want the figure more accurately than "between $.85 and $.95," as given above, we proceed as follows:—The 512th man from the bottom is the 82nd man in the group between $.85 and $.95, for there are 430 earning less than $.85; this group contains 169; if they were distributed regularly, 17 to each cent, the 82nd man would be half-way through this group, between $.89 and $.90. The hypothesis of even distribution is sufficiently correct for most purposes, and this method affords a sufficiently accurate means of determining the wage of the workers at the tenth places. The resulting figures are given on p. 92. If, however, we want to know the wage of the half-way man more exactly, we see from the half-dollar groups that it is between $1.25 and $1.75, a rough approximation shows it to lie probably between $1.45 and $1.55, and then we rapidly turn through our original data, isolating the wages at $1.46, $1.47, . . . $1.55.*

A slight modification of this method is also useful. Take the average of the lowest 512 (or tenth), namely, $.70½; of the next, namely, $1.03; and so on (see p. 92). These figures also give a vivid view, and are very convenient for comparisons with other groups.

The figures so far apply to only half of the data in the Senate Report. On p. 92 the whole are tabulated to give the average wages of the successive tenths. A comparison of the two groups so obtained shows how far the first half was typical of the whole. This method will be dealt with in a later chapter.

* On this method see pages 127, 128.

TABULATION OF WAGES—AMERICAN FIGURES, 1891.

1. Earning Daily Wages.		2. No. of Persons.	3.		4. No. of Persons.	5.		6. No. of Persons.	7. Percentage.	8. Average Wage in Group.	9.
$ as much as	and less than		$ as much as	and less than		$ as much as	and less than				
.25	.35	1	.25	.45	16						
.35	.45	15									$ instead $
.45	.55	59	.45	.65	144	.25	.75	317	6.2	.62 of	.50
.55	.65	85									
.65	.75	157	.65	.85	270						
.75	.85	113									
.85	.95	169	.85	1.05	370	.75	1.25	1,472	28.7	1.09	1.00
.95	1.05	201									
1.05	1.15	304	1.05	1.25	989						
1.15	1.25	685									
1.25	1.35	99	1.25	1.45	557						
1.35	1.45	458									
1.45	1.55	466	1.45	1.65	538	1.25	1.75	1,297	25.3	1.49	1.50
1.55	1.65	72									
1.65	1.75	202	1.65	1.85	531						
1.75	1.85	329									
1.85	1.95	58	1.85	2.05	331	1.75	2.25	970	18.9	1.99	2.00
1.95	2.05	273									
2.05	2.15	45	2.05	2.25	310						
2.15	2.25	265									
2.25	2.35	33	2.25	2.45	134						
2.35	2.45	101									
2.45	2.55	196	2.45	2.65	209	2.25	2.75	506	9.9	2.53	2.50
2.55	26.5	13									
2.65	2.75	163	2.65	2.85	165						
2.75	2.85	2									
2.85	2.95	15	2.85	3.05	144	2.75	3.25	198	3.9	3.04	3.00
2.95	3.05	129									
3.05	3.15	5	3.05	3.25	52						
3.15	3.25	47									
3.25	3.35	12	3.25	3.45	12						
3.35	3.45	0									
3.45	3.55	221	3.45	3.65	226	3.25	3.75	254	5.0	3.51	3.50
3.55	3.65	5									
3.65	3.75	16	3.65	3.85	27						
3.75	3.85	11									
3.85	3.95	0	3.85	4.05	82	3.75	4.25	96	1.9	4.00	4.00
3.95	4.05	82									
4.05	4.15	0	4.05	4.25	3						
4.15	4.25	3									
4.25	4.35	0	4.25	4.45	0						
4.35	4.45	0									
4.45	4.55	3	4.45	4.65	4	4.25	4.75	4	0	4.50	4.50
4.55	4.65	1									
4.65	4.75	0	4.65	4.85	0						
4.75	4.85	0									
4.85	4.95	0	4.85	5.05	8	4.75	5.25	8	.2	5.00	5.00
4.95	5.05	8									
5.05	5.15	0	5.05	5.25	0						
5.15	5.25	0					At 5.35	1		5.35	5.25
5.25	5.35	1	5.25	5.35	1						
Totals		5,123			5,123			5,123	100		
Average Wage		$1.731							Average Wage		$1.70

Wages of "Tenth" Men (*deciles*).	
Lowest Wage	$.25
Wages of Men: $\frac{1}{10}$th up Group	.89
$\frac{2}{10}$th ,,	1.12
$\frac{3}{10}$th ,,	1.22
$\frac{4}{10}$th ,,	1.39
$\frac{5}{10}$th ,,	1.49
$\frac{6}{10}$th ,,	1.75
$\frac{7}{10}$th ,,	1.99
$\frac{8}{10}$th ,,	2.36
$\frac{9}{10}$th ,,	2.98
Highest Wage	5.35

Average Wage of		Same for 10,000 Workers.
Lowest tenth	$.70	.79
Second ,,	1.03	1.00
Third ,,	1.18	1.24
Fourth ,,	1.28	1.50
Fifth ,,	1.44	1.50
Sixth ,,	1.59	1.88
Seventh ,,	1.86	2.00
Eighth ,,	2.14	2.22
Ninth ,,	2.59	2.58
Highest ,,	3.51	3.55
General Average	1.731	1.82

The tabulation of the data collected for the WAGE CENSUS on such forms as that on p. 36, illustrates well some of the difficulties involved. The items given on the main part of the schedule are of this kind :—

	No.	Average Wage.	
Spinners—Time :	6	: 12s.	: 56½ hours.

Tabulation in the wage census.

Such returns are not perfectly definite, for if many are employed in the same occupation in a mill, it is possible that they will earn at different rates. Thus this entry of 6 at 12s. might arise from either 6 men each earning 12s., or 2 at 10s., 2 at 12s., 2 at 14s. (average 12s.); or 4 at 12s., 1 at 15s., 1 at 11s.; or 5 at 12s. and 1 at 18s.—12s. being the general rate, but not the average, in these last two alternatives. Since the purpose of the wage census was to give a comprehensive account of wages adapted for use in all investigations, it should show the numbers in all trades and subdivisions of employment by age, sex, and district, the average and general rate of pay for each group, and sufficient details to show the distribution about the average in each group, for a mere average may conceal exceptionally high or exceptionally low wages.

On inquiry at the Labour Department as to whether the original information had been given in a more detailed form than the line above, or whether divergencies might be concealed, the author learnt that the subdivision of occupations had been carried to such an extent, that in practice, where there was any great variation in the wages of workers under one heading, that head-

ing had been split up, so that each group was separately entered, or that several groups were distinguished under one heading; and that when there was reason to believe from the light of other returns that this had not been done, supplementary inquiries were made on this point, so that the original data were detailed enough for any requisite fineness of tabulation.

The problem then was to tabulate the answers from the various factories in a district, to show clearly and succinctly the distribution of wages in each subdivision and in the whole. It can hardly be said with confidence that the method adopted, of which a specimen is given on p. 94, is entirely satisfactory.

To clear our ideas let us suppose that the details on which the line relating to throwsters (time) was based were as follows:—

3	earning	14/	- "average minimum rate."
14	„	15/	
6	„	15/6	
20	„	16/	68 within 10 per cent. of the average for all, which is 17/7.
10	„	17/6	
20	„	18/	
8	„	18/6	
10	„	19/	
10	„	20/6	18 earning 20/11 on the average.
8	„	21/5	

The process adopted in the tabulation may be supposed to have been to separate from the whole group of returns a small group of old men or inferior workers earning far below the average, and enter them as a distinct minimum group, and to separate a small group of the most skilled workers and enter them as a maximum group. This is better than giving simply the highest and lowest of the individual wages, for either of these may be due to exceptional circumstances, and may be quite a long way from that paid to any other person. The exact size of these extreme groups must be determined from inspection of the returns themselves. After this has been done, the remaining wages may not be grouped close together; in the example taken they are scattered between 15s. and 19s. To give some clue as to this distribution the number earning within 10 per cent. of the

Various methods possible.

RATES OF WAGES.*

SILK MANUFACTURE.

Average Rates of Wages of Operatives Employed in the Silk Manufacture in Various Districts, at 1st October 1886.

DISTRICT I.—CHESHIRE, STAFFS, NOTTS, DERBY, AND WARWICK.

Total Number of Persons employed at 1st October 1886, as per Returns, 5,279.

Total amount paid in Wages in 1885, as per Returns, £129,588.

OCCUPATIONS.	Number employed as per Returns.	Percentage that the Number returned under each Occupation is of Total Number of Persons returned.	AVERAGE Weekly Rate of Wages.	Number returned at Rates within 10 per cent above and below the average.	Average *Maximum* Weekly Rate of Wages.	Number on which Average Maximum Rate is based.	Average *Minimum* Weekly Rate of Wages.	Number on which Average Minimum Rate is based.
	MEN.							
	Total Number, 1,085.				Hours of Labour per week, 56½.			
THROWN SILK—			*s. d.*		*s. d.*		*s. d.*	
Overlookers and Stewards - - - Time	79	1.5	23 3	35	31 3	4	20 0	29
Assistant Overlookers - - - Time	9	.2	17 3	6	19 0	...	15 0	...
Spinners - - - - - - Time	24	.5	11 9	15	13 0	5	9 3	4
Millmen (Throwsters) - - - Time	109	2.1	17 7	68	20 11	18	14 0	3
Millmen (Throwsters) - - - Piece	9	.2	18 8	7	...	...	...	...

* From *Wages in the Minor Textile Trades* (C.—6161 of 1890).

average is stated; this is probably the best way if only one column can be devoted to it, but 10 per cent. is a wide limit to adopt. Another method would be to give the limits within which the wages of the 10 per cent. of the earners above and 10 per cent. below the average were contained: in this case 16s. and 18s.

If, however, not more than 8 columns are to be devoted to each group, the following arrangement would give much more definite information, and it could have been made from the data in hand, and would be well adapted for all the purposes for which it would be required.

Number employed	109
General average	17/7
Average of lowest tenth *	14/9
Quartile †	16/
Median †	18/
Quartile †	19/
Average of highest tenth *	21/2

The general summary.

We are fortunately not dependent solely on the tabulation as given above, for wages in industries as a whole are also tabulated on the following plan, which is in a form most useful for purposes of comparison (p. 96).

The lines giving percentages are most useful. We can at a glance compare the levels of wages in different industries. Thus in the cotton manufacture the average wage is 2s. higher than in the woollen; and in the cotton there is a large group of highly skilled workers earning from 30s. to 35s., while in the woollen nearly half are close to the average, earning between 20s. and 25s. In the jute and linen manufactures the averages are nearly the same, but in the former a larger proportion are below the 15s. limit. In the silk manufacture there is an aristocracy as in the cotton, but it is smaller and better paid, for 12 per cent. earn more than 35s. This table is a masterpiece of concentration and clearness.

We will discuss next the tabulation of the figures relating

* *Vide* p. 92. † *Vide* p. 124.

NUMBER AND PERCENTAGE OF PERSONS EMPLOYED AT VARIOUS RATES OF WAGES.*

Table showing the average Normal Wages paid to men in the undermentioned employments, and the Number and Proportion of men paid at different rates, at October 1886.

		Under 10s.	Of 10s. and under 15s.	Of 15s. and under 20s.	Of 20s. and under 25s.	Of 25s. and under 30s.	Of 30s. and under 35s.	Of 35s. and under 40s.	Above 40s.	TOTAL.	Average Wages per Head. s.	d.
Cotton Manufacture	Number	2	370	8,793	8,822	4,525	7,283	1,582	812	32,189	25	3
	Per cent.	...	*1.2*	*27.3*	*27.4*	*14.1*	*22.6*	*4.9*	*2.5*	*100*		
Woollen ,,	Number	...	146	3,377	5,559	1,725	705	392	344	12,248	23	2
	Per cent.	...	*1.2*	*27.6*	*45.4*	*14.1*	*5.7*	*3.2*	*2.8*	*100*		
Worsted and Stuff Manufacture	Number	...	835	1,705	909	2,635	879	28	14	7,005	23	4
	Per cent.	...	*11.9*	*24.3*	*13.0*	*37.6*	*12.6*	*0.4*	*0.2*	*100*		
Linen Manufacture	Number	192	780	2,952	2,070	416	290	39	68	6,807	19	9
	Per cent.	*2.8*	*11.4*	*43.4*	*30.4*	*6.1*	*4.3*	*0.6*	*1.0*	*100*		
Jute ,,	Number	...	565	1,038	964	127	53	52	...	2,799	19	4
	Per cent.	...	*20.2*	*37.1*	*34.4*	*4.5*	*1.9*	*1.9*	...	*100*		
Hemp ,,	Number	...	25	300	581	168	39	94	25	1,232	23	6
	Per cent.	...	*2.0*	*24.4*	*47.2*	*13.6*	*3.2*	*7.6*	*2.0*	*100*		
Silk ,,	Number	...	324	881	367	278	121	273	4	2,248	22	3
	Per cent.	...	*14.4*	*39.2*	*16.3*	*12.4*	*5.4*	*12.1*	*0.2*	*100*		
Carpet ,,	Number	...	...	130	183	834	100	15	30	1,292	26	7
	Per cent.	...	...	*10.1*	*14.2*	*64.5*	*7.7*	*1.2*	*2.3*	*100*		
Hosiery ,,	Number	...	...	296	458	51	190	75	...	1,070	24	5
	Per cent.	...	...	*27.7*	*42.8*	*4.8*	*17.7*	*7.0*	...	*100*		

* *General Report on Wages* (C.—6889 of 1893).

to CHANGES in RATES of WAGES collected by the Labour Department. Specimens of the forms by which such information is obtained were given among those relating to strikes (p. 57). Referring to them, it will be seen that the facts given are the occupations and numbers affected, the dates from which the changes took place, and the wages and hours in a full week exclusive of overtime (a definition corresponding exactly to that used for the wage census) before and after the change.

Tabulation of change of wages returns.

EXTRACT FROM TABLE showing the Changes in Rates of Wages and Hours of Labour of Ordinary Agricultural Labourers in Various Districts of the United Kingdom in 1894, so far as reported to the Board of Trade.*

County and Union.	Particulars of Changes in Summer Wages. (1894 compared with 1893.)		Particulars of Changes in Winter Wages. (1894 compared with 1893.)		No. of Male Agricultural Labourers, Farm Servants, Shepherds, Horsekeepers, Horsemen, Teamsters, Carters, in '91.
	Increase.	Decrease.	Increase.	Decrease.	
		Per Week.	Per Week.	Per Week.	
LINCOLNSHIRE—					
Gainsborough	...	...	...	1/6(15/to13/6)	2,466
Louth	...	...	...	1/6(13/6to12/)	3,932
Spilsby	...	...	...	1/6(13/6to12/)	3,288
NORFOLK—					
Aylsham	...	1/(12/to11/)	...	...	2,576
Docking	...	6d.(12/6to12/)	1/(10/-11/)	...	2,487
Flegg, East and West	...	1/(12/to11/)	...	1/(11/to10/)	1,108
Forehoe	...	...	...	1/(11/to10/)	1,448

* From the second *Annual Report on Changes of Wages*, pp. 198-9; a little compressed.

EXTRACTS FROM TABLE showing the Changes in Rates of Wages of Ordinary Agricultural Labourers in Various Districts of the United Kingdom in the Summer of 1895, so far as reported to the Board of Trade.*

County and Union.	No. of Male Agricultural Labourers, Farm Servants, Shepherds, Horsekeepers, Horsemen, Teamsters, Carters, in 1891.	Particulars of Changes in Summer Wages (1895 compared with 1894). *Decreases in italics.*	Weekly Rate of Wages in Summer. 1894.	Weekly Rate of Wages in Summer. 1895.
		Per Week.	s. d.	s. d.
DURHAM—				
Stockton*	437	*Decrease of 6d.*	17 6	17 0
Teesdale (Barnard Castle Rural Dist.).*	669†	Advance of 6d.	17 6	18 0
OXFORDSHIRE—				
Headington	1,118	*Decrease of 1s.*	12 0	11 0
Henley (Hambleden Rural Dist., Bucks).	1,587†	*Decrease of 1s.*	12 0 to 14 0	11 0 to 13 0
NORFOLK—				
Flegg, East & West	1,108	*Decrease of 1s.*	11 0	10 0
Forehoe	1,448	*Decrease of 1s.*	11 0	10 0
Henstead	1,504	*Decrease of 1s.*	11 0	10 0
Mitford and Launditch	3,622	*Decrease of 1s.*	11 0	10 0
Smallburgh	2,264‡	*Decrease of 1s.*	11 0	10 0
Swaffham	1,942	*Decrease of 1s.*	11 0	10 0
Wayland	1,535	*Decrease of 1s.*	11 0	10 0
CARNARVONSHIRE—				
Carnarvon (Gwyrfai Rural Dist.).	1,124†	Labourers without food, advance of 1s.	19 0	20 0
		Labourers with food, advance of 1s.	11 0	12 0

* Agricultural labourers in this district are hired in March and April for a year certain, and the change noted applies to the whole year, and not to the summer only.

† The number of agricultural labourers, &c., is for the Poor Law Union, but the change applies to the Rural District only.

‡ This number is partly estimated.

---

*From the third *Annual Report on Changes of Wages*, pp. 118, 119, 121 (typography adapted).

The adjoining tables give examples of the way in which the changes in agricultural wages were tabulated in the Second and Third Report on Changes in Rates of Wages and Hours of Labour. In the first table space is wasted by devoting separate columns to increases and decreases, with the intention of making the table distinct; while it is not clear whether "Winter 1894" means the winter beginning in or that ending in that year.

Agricultural wages: Change in tabulation.

In the second table, which refers to summer wages only, the columns are rearranged; and increases and decreases printed in the same column, the latter in italics. In the Fifth Report all the information is printed in a clearer way, thus :—

WINTER WAGES.*

District.	Number.	Weekly Rates.		Increase or Decrease per Week in 1897.	
		Jan. '96.	Jan. '97.	Increase.	Decrease.
		s. d.	*s. d.*	*s. d.*	
Tendring	3,113	10 0	11 0	1 0	...

The tabulation is repeated for the summer.

The weakness in these agricultural returns is in the numbers column. In the returns from other industries the numbers given are those actually affected, but in this case it is not found possible to obtain this number correctly, and the number entered is that found under "agricultural labourers" in the 1891 census, which includes the various categories as given in the above table. When a change of wages takes place in a rural district, we may perhaps assume that it is likely to be general, though, if it was a reduction, it might not be made by the better employers; and though the change will not take place in the same week throughout the district, there is not likely to be much variation in this respect. The change is generally made at the time that winter wages give place to summer, or summer to winter; and a slight increase or decrease may take place by making the winter reduction or the summer advance later than usual. On the whole, little error will be introduced by assuming that the change stated affects all the adult agricultural labourers in the district,

The number affected.

* From the fifth *Annual Report on Changes of Wages*, p. 145.

and it is quite probable that a proportional change* will take place in the wages of horsekeepers, shepherds, and others, though it may not in the case of boys, or old men who are earning less than the district rate. The question, "Approximate number of able-bodied labourers in parish?" is asked on the inquiry form, but as the answers are not used, it may be assumed that they are generally not given with sufficient exactness.

Lack of data.

The object of the whole tabulation is to show the change in the national weekly wages bill, but many details are lacking for the complete calculation. In the case of agricultural labourers, we need, in addition to these data, accurate statements of the change of additional earnings, special payments, and payment in kinds. In all cases we need a more complete account of the whole wage-bill as well as the change. For agricultural labourers the material has just been published by the Labour Department;* every year it receives returns from most of the 600 unions as to wages at all seasons, whether there has been a change or not.

Changes in county rates.

The looseness in the returns as to numbers does not prevent our calculating the change in the county or country rates, for the numbers in each district affected by the change may be expected to bear the same proportion to the numbers given in the census returns, as the number of agricultural labourers of the same class in the whole county or country does to the census number.

The calculation for Durham in the above table for the changes in summer wages 1894-95 may be performed as follows:—

	Average before change.	Change.	Proportional number affected.	Amount of change on wage-bill.
	s. d.			*s. d.*
Stockton - -	17 6	−6d.	4	−2 0
Teesdale - -	17 6	+6d.	7	+3 6

Total change in county, + 1s. 6d.
Proportional number in county, 73.

Effect on county average, $\frac{1/6}{73} = \frac{1}{4}$d.

Here, for simplicity of calculation, the numbers affected are

* On these points see Mr Wilson Fox's Report on *Wages and Earnings of Agricultural Labourers*, 1900, p. 50, and pp. 111-157.

taken to the nearest 100, a process which is not likely to affect the average perceptibly.* This rough method is likely to give the result as accurately as the original data make possible. A similar process with suitable modifications can be applied to the changes tabulated for other industries. The summary of such returns for agriculture for all counties is as follows :—

COMPARISON OF THE NET EFFECT OF THE CHANGES OF CASH WAGES per Week paid in the Years 1896 and 1895 in certain Districts in England and Wales.†

DISTRICT.	WAGES IN 1896 AS COMPARED WITH 1895.			WAGES IN 1895 AS COMPARED WITH 1894.		
	Total ** Number.	Net Effect of Changes on Weekly Wages. Increase (+) and Decrease (−).		Total ** Number.	Net Effect of Changes on Weekly Wages. Increase (+) and Decrease (−).	
		Total.	Per Head.		Total.	Per Head.
ENGLAND—		£	s. d.		£	d.
Northern Counties -	5,662	−43	−0 1¾	3,766	+44	+2¾
Yorkshire, Lancashire, and Cheshire	2,897	+100	+0 8¼	3,942	−126	−7¾
Eastern and Midland Counties - -	69,869	+666	+0 2¼	89,576	−2,045	−5½
Southern and Western Counties -	20,901	−340	−0 4	20,441	−575	−6¾
WALES - - -	...	...	...	2,165	+73	+8¼
Total - - -	99,329	+383	+0 1	119,890	−2,629	−5¼

** The number given is the total of male agricultural labourers, farm servants, shepherds, horse-keepers, in 1891, in the Poor Law Unions in which the changes took place.

* The corresponding calculations for Oxfordshire are :—

12/	−1/	11	−11/
13/	−1/	16	−16/
			−27/

Effect on county average, $\frac{-27/}{161} = -2d.$

For Norfolk :—

12/	−1/	134	−134/

Effect on county average, $\frac{-134/}{425} = -4d.$

† From the fourth *Annual Report on Changes of Wages*, p. xliv.

29224

The value of this table is not obvious. It seems of little importance to know how many persons were affected altogether; though it is of some value to learn from a previous table that 58,578 persons received increases, and 40,751 decreases in 1896. This total of persons affected is constantly given in these tables; if a person receives an increase of 1s. one month, and loses it the next, he is counted as 2, and his contribution to the next column (net effect of change) is zero. This −£43 may mean that 2,000 persons received a decrease of 1s. each, and the remaining 3,662 (same or different persons) an increase of 3¾d. each, or any other figures which would give the same total. The change per head in the next column is unimportant; it only shows an arithmetical quotient with no concrete meaning that can be expressed in words. If it was replaced by another quotient, viz., $\frac{£43}{n}$, where n is the number of agricultural labourers in the Northern Counties, we should know the effect on average wages. In fact, the table would be more useful thus:—

Criticism of summary table.

APPROXIMATE EFFECT OF CHANGES ON NATIONAL WEEKLY WAGE BILL.

District.	Increases.		Decreases.		Net Change.	Total No. Employed.	Average Change.
	No. affected.	Total.	No. affected.	Total.			

The figures given supply an example of the common practice of carrying out into detail a calculation which depends originally on incorrect numbers, in this case the number employed, and is therefore misleading throughout. Till the average (useless here in any case) is taken, the error in this quantity has no injurious effect. As shown above, the average here given could be replaced by another which would be of use, and which would be correct within limits that could be defined, and would be narrow enough for most purposes.

Further, since the column of numbers affected is admittedly wrong, the figures should be given to the nearest 1,000 rather

than to units, even if no attempt was made to estimate the new figure; "between 5,000 and 6,000 are affected" is a more useful and correct statement than "5,662 persons belonged in 1891 to a class in some undefined way connected with that in question in 1896."

The discussion of Group C, the tabulation of non-numerical answers, must be postponed till we have analysed the nature and use of averages.

CHAPTER V.

AVERAGES.

CHAPTER V.

AVERAGES.

IT is natural, in a book with the present title, to allot a considerable space to averages. By the use of averages complex groups and large numbers are presented in a few significant words or figures; and thus the two definitions of statistics, the *Science of Averages* and the *Science of Large Numbers*, are reconciled.

Some writers have attempted to draw a distinction between *averages* and *means*, but no general agreement has been reached as to the exact senses in which the words are to be separately applied.* The best distinction may be made by deciding that an average is a purely arithmetical conception, such as the average length of life in a varied population, which does not correspond to any particular group, but is only a short way of expressing an arithmetical result; while the word "mean" is to be applied to some objective quantity, such as the mean height of Englishmen, about which all height-measurements are grouped according to a definite law, which will be discussed in the sequel.†

Averages and means.

A. ARITHMETIC AVERAGES.—We may rapidly pass by some of the common uses of the word "average," and pick out those which will prove of use in statistics. An average is sometimes used merely to save big figures. The average weight of the University crew is given, only because it is more usual to speak of a man's weight being 12½ stone than of eight men's weight being 12½ cwt., and it is easier to connect the former with men's weight in general. Similarly, if we are comparing the value of the exportations of some commodity in two periods

---

* Compare the article "Moyenne," by Dr Bertillon, in *Dictionaire encyclopédique des Sciences Médicales*, with this chapter. See also the paper by Dr Venn in the *Statistical Journal*, 1891, and chap. xviii. in his *Logic of Chance*.

† See Part II., Section 1, *infra*.

of ten years each, we should say that the yearly average in the period 1870-79 was £10,000,000, and in 1880-89 was £11,000,000, rather than that the totals were £100,000,000 and £110,000,000. This leads to the second ordinary use of the word. If we were comparing the ten years 1870-79 with the eleven years 1880-90, and the totals in the periods were £100,000,000 and £132,000,000 respectively, we should obtain no grasp of the difference till we had reduced them to a common denominator by dividing by the number of years, and found that the averages in the two periods were £10,000,000 and £12,000,000. This class of averages is well known in cricket; sometimes the total number of runs made or wickets taken by each cricketer are stated also, but these are rather as so-called statistical curiosities than as having much bearing on the skill or luck of the players. The numbers by which the seasons' performances are judged are the quotients of the number of runs by the number of innings, of the number of wickets by the number of runs, and so on, all quantities being reduced to a common denominator. A consideration of the best methods of comparing cricketers or counties, and an exposure of the fallacies inherent in the present system, would afford a useful exercise in the use of averages and the choice of the most appropriate kind. The average in this sense is very common in mechanics. The average pressure per square inch, the average work done by an engine per minute, the average speed of a train, are quantities which it is frequently necessary to use. Such an expression as the average rate of interest is precisely similar.

The common denominator.

It will be clear that percentage is a special case of this use of average. It is useless when comparing the growths of population or of trade to give only the whole numbers. An increase of 50,000 in the population of London is not so significant as one of 10,000 in that of Harrow; they must be expressed as increases of 1 per cent. and 150 per cent., say, before their meaning can be appreciated, and this is the same thing as giving the average increase to 100 inhabitants. For this reason the records of births, deaths, and marriages are always given as rates—so many per 1,000 inhabitants; and in these cases a double average is given, for the rates signify so many per 1,000 inhabitants per annum.

Averages as rates.

Another extension of the same use is found when quantities are reduced to rates "per head" of the population. This use is solely for comparison, and the principle employed is that of the common denominator. It would be futile to state that the amount spent on drink was, say, £100,000,000 in 1860 and £110,000,000 in 1890; but the corresponding statements that the amounts were £3. 10s. per head in 1860 and £2. 15s. per head in 1890 would make a comparison possible. Or, to take a better instance: in studying the increments in the values of England's foreign trade, an entirely wrong view is obtained, unless we calculate for each year the value per head of the population, instead of looking only at the totals. A neglect of this division would make municipal expenditure appear to be growing much faster than it really is; and in preparing any comparative summary of figures, it is always necessary to consider whether such an average should be taken.

Preliminary definition.

So far, the averages considered are simply arithmetical, and satisfy the following definition:—

Average × *number to which it applies* = *total quantity dealt with.*

e.g. Average weight × number of crew = total weight of crew.

Average value of imports per head of population × number of population = total value of imports, and so on.

Its inapplicability.

The following question, however, will lead us further. The average weekly agricultural wages in 1892 in Wilts, Dorset, Devon, Cornwall, and Somerset were 10s., 10s., 13s. 6d., 14s., 11s. respectively. What was the average in the south-west of England?

The simplest method is to say, the average was

$$\frac{10s. + 10s. + 13s.\ 6d. + 14s. + 11s.}{5} = \frac{58s.\ 6d.}{5} = 11s.\ 8.4d.$$

and for many purposes this would be sufficient; but it does not satisfy the above definition. For when we ask the double question "11s. 8.4d. multiplied by what number equals what total?", we can only answer that 11s. 8.4d. multiplied by the number of items equals the sum of items.

We must consider further what we understand by the expressions "average wage in each county," and "average wage in the group of five counties."

It may be supposed that the average wage in Wilts, for instance, was compiled by getting returns from different villages, say 12s., 11s., 9s., 9s. 6d., 10s. 6d., 9s., 9s., adding them and dividing by the number of villages. This of course satisfies our definition no better than the former. What is to be understood by the average in each village? If our present definition is to be satisfied, it should be the total of the wages paid in the village divided by the number of workers. It is hardly necessary to say that this total is never found in such an investigation, and the average is given from observation or by guess-work, not by calculation.

If, however, the village average was correct, and we had returns from all the villages in the county, we should find the county average as follows:—

$$\frac{12/\times 200+11/\times 150+9/\times 300+9/6\times 150+10/6\times 400+9/\times 200+9/\times 200}{200+150+300+150+400+200+200}=9/11.8,$$

where the numbers in the denominator are the numbers of labourers in the respective villages. We should then have the same result as if we had had the wages of all the labourers in the county put down on a sheet, added up, and divided by their number, and the average would satisfy the definition.

It is clear that we can simplify this arithmetical work, for if we divide throughout by 50 we get the same result; this is as if we said there were 4, 3, 6 . . . labourers in the villages instead of 200, 150, . . . Thus we get the same result if we take numbers proportional to the total numbers of the labourers instead of the actual numbers. This plan has two advantages: first, that though we do not know the numbers of labourers, we know numbers nearly proportional to them, viz., those included in the census returns under the general headings relating to agriculture; and secondly, we need not choose our numbers with absolute exactness; thus the numbers of labourers above given may be supposed to be round numbers substituted for 213, 145, 320 . . .; and it will presently be seen that such differences hardly affect the average. We idealize the village, and suppose it to contain round numbers; and then for the numerical work take simple numbers proportional to these. This is important as simplifying numerical work.

Averages obtained for the county in this way do not ab-

solutely satisfy our definition, but are very nearly equal to those that do. We can then proceed to take the average for the south-west of England on the same principles.

B. WEIGHTED AVERAGES.—This discussion introduces and gives an example of the very important statistical method known as "weighting the average." We may illustrate it further from the same figures by considering what weights to apply to get this average for South-West England. We may find the number of agricultural labourers in the counties and work out the average thus: $\frac{10s. \times 20{,}000 + 10s. \times 30{,}000 + \ldots}{20{,}000 + 30{,}000 + \ldots}$; or we may argue that since we have no means of knowing the exact numbers of labourers we may as well arrange the weights, according to the importance of the counties, say 20,000, 30,000, &c., from some other point of view, and take numbers representing such quantities as the amounts of wheat produced, the area, or the rate of increase of population. In this particular case these methods would be absurd, but in other problems the weights are not so obvious. Suppose, for example, that we are considering the attraction of London on the inhabitants of various counties; that we are told that so many immigrants arrive from Essex, Norfolk, and Suffolk, and so many from Stafford and Worcester, and we are asked to compare the attractive power on the agricultural and manufacturing counties. Should we weight the numbers given by the total numbers of inhabitants of the contributing counties, or by their distance from London, or by some quantity derived from these?

A more practical problem, the classical and most useful application of weights, is the formation of an index number for the change of prices by fitting-suitable weights to the changes measured in the prices of various commodities. This will be considered separately,* but it is best to deal with the first principles here. It is required to find the change in the value of gold when measured by the prices of other commodities. Suppose that we are given that the prices of certain commodities between two years were in the following ratios:—

Index-numbers.

* See *infra*, Chap. IX

	Wheat.	Silver.	Meat.	Sugar.	Cotton.
First Year - - -	100	100	100	100	100
Second Year - -	77	60	90	40	85

The simplest way to estimate for the general fall in price is to take the simple average of the numbers in the second year, viz., 70.4; and say that general prices in the second year were 70.4 per cent. of those in the first, and the value of gold had increased in the ratio 70.4 : 100 when expressed in commodities. But it is at once clear that we cannot allow the commodities given to have equal influences on the result; wheat is of greater importance than sugar and meat than silver; and again we have taken arbitrarily three items to represent food and one for clothing; we need some means of deciding relative importance. Suppose we decide that wheat, cotton, meat, and sugar are respectively 7, 4, 3 times and twice as important as silver, we should get the following table:—

Commodity.	Relative Price in Second Year.	Weight Assigned.	Product.
Wheat - - - -	77	7	539
Silver - - - -	60	1	60
Meat - - - -	90	3	270
Sugar - - - -	40	2	80
Cotton - - -	85	4	340
	352	17	1289

Weighted average is $\frac{1289}{17} = 75.8$

Unweighted average $\frac{352}{5} = 70.4$

This process is equivalent to writing down the price of wheat seven times, silver once, meat thrice, &c., and then taking the simple average of these numbers.

The idea is made clearer by the mechanical analogy in which the word *weight* originated. Suppose a uniform weightless rigid rod graduated in 100 equal divisions, and equal weights hung at the 77th, 60th, 90th, 40th, and 85th divisions from one end; the rod will then balance at a point corresponding to the unweighted average, 70.4 intervals

Mechanical illustration.

from the same end. Now, suppose the equa. weights replaced by weights of 7, 1, 3, 2, 4 lbs. respectively, and the rod will balance at a point corresponding to the weighted averages 75.8 intervals from the same end. The further any particular mass is moved, or the heavier it is, the more the centre of gravity will be shifted; and this clearly corresponds to the influence we should wish the various prices to have in the statistical problem. The formula in use in Statics, $\bar{x} = \frac{Smx}{Sm}$, which corresponds to the arithmetic on the previous page, can also be used in Statistics.

The discussion of the proper weights to be used in this and other averages has occupied a space in statistical literature out of all proportion to its significance, for it may be said at once that no great importance need be attached to the special choice of weights; one of the most convenient facts of statistical theory is that, given certain conditions, the same result is obtained whatever logical system of weights is applied. We must postpone the mathematical analysis of this proposition, but may offer immediately some arithmetical illustrations.

The small effect of weights.

The table on the next page affords an example of this principle,* and is worth careful study. At the commencement of the Wage Census, circulars were sent to all the principal firms in all well-located trades, asking for details as to wages. Of these some were not returned, and the numbers allotted in the Final Report to each trade are not the numbers which actually belong to the trade in the whole country, but the numbers of those in the firms which made returns. The average wage given is not therefore the arithmetic average for these trades for the whole country corresponding to the definition given above for average, but the average of the average wages as returned in each trade weighted by the numbers for whom returns were made; so that the average wage given for the whole group of trades might have proved to be different, if with the same average in each trade the returns had been complete. It is very unlikely, however, that there would have been any great difference. In the table several systems of weighting are used; the first are the numbers in these returns, giving an average, 24s. 7d.; the second are the numbers be-

Example from the Wage Census.

* From the *Statistical Journal*, December 1897, with corrections.

Examples of the Smallness of the Change Introduced by Difference in Systems of Weighting.

From the Wage Census.						
Trade.	Average Wages (Men).	Number Included in Returns	Numbers Employed in Trade when known. Unit 1,000	Arbitrary System of Weights.	Equal Weights.
	s. *d.*				
Cotton Manufacture	25 3	32,189	142	144	1
Woollen "	23 2	12,248	54	172	1
Worsted and Stuff Manufacture	23 4	7,005	38	219	1
Linen Manufacture	19 9	6,807	22	96	1
Jute "	19 4	2,799	9	23	1
Hemp, &c., "	23 6	1,232	3	78	1
Silk "	22 3	2,248	10	189	1
Carpet "	26 7	1,292	0	213	1
Hosiery "	24 5	1,070	8	287	1
Lace "	27 3	593	8	51	1
Smallwares "	20 2	2,734	0	225	1
Flock and Shoddy Manufacture	21 2	330	2	200	1
Coal, Iron Ore, and Ironstone Mines	22 11	67,429	57	142	1
Metalliferous Mines	16 6	5,046	0	190	1
Shale Mines and Paraffin Oil Works	25 0	3,021	0	207	1
Slate Mines and Quarries	22 1	6,933	12	232	1
Granite Quarries and Works	21 11	2,315		206	1
Stone Quarries	23 10	3,956		34	1
China, Clay, &c., Works	18 8	499	0	3	1
Police	27 7	52,682	58	224	1
Roads, Pavements, and Sewers	20 9	24,276	0	29	1
Gasworks	27 2	27,965	0	40	1
Waterworks	24 9	5,187	0	151	1
Pig Iron (Blast Furnaces)	24 6	6,234	0	128	1
General Engineering Iron and Brass Foundries and Machinery Trades	25 9	41,658	200	173	1
Shipbuilding, Iron and Steel	29 3	10,661	80	228	1
Tinplate Works	33 5	11,514	0	178	1
Saw Mills	24 3	2,088	0	174	1
Brass Works and Metal Wares	29 7	1,838	0	222	1
Shipbuilding, Wood	28 4	454	0	79	1
Cooperage Works	30 5	327	0	165	1
Coach and Carriage Building	26 6	1,664	0	28	1
Boot and Shoe Making	24 3	2,902	0	142	1
Breweries	24 3	8,366	0	46	1
Distilleries	20 4	1,795	0	129	1
Brick and Tile, &c., Making	22 10	3,188	0	55	1
Chemical Manure Works	23 0	1,054	0	210	1
Railway Carriage and Wagon Building	25 2	2,239	0	233	1
Averages	...	*s.* *d.* 24 7	*s.* *d.* 25 3	*s.* *d.* 24 5¼	*s.* *d.* 24 2

longing to each trade according to the census when they are above a certain minimum, giving an average 25s. 3d.; the third is a purely arbitrary list of figures taken from a source which has no connection with wages, and the average is 24s. 5¼d.; the last is the unweighted average, that is, all the weights are equal, and the average is now 24s. 2d. These averages are close together, while the original items vary from 16s. 6d. to 30s. 5d. It is to be noticed that the true weights are not known in this case, but that owing to this principle we are able to dispense with them entirely.

Uniformity of average under many systems of weights.

The problem dealt with in the next table is to find the average weekly agricultural wage in England and Wales from the returns for Michaelmas 1869 and Lady Day 1870, given in columns 1 and 2. There are very many different ways of taking this average, some of which are as follows:—Take the average of summer and autumn for each county, as in column 3, and then the unweighted average of these 45 numbers; this is 12s. 7d. Suppose the summer wage to be paid twice as long as the autumn wage, as in column 4, and proceed as before; the average is 12s. 5½d., the slight difference being due to the inclusion of harvest payments in the Michaelmas wage, which makes them higher on the whole than the summer wages. Again, divide the counties into geographical groups, take the simple average for each group (the figures marked *a* in column 3 and *b* in column 4) and weight these by the figures marked *c* in column 5, the numbers of agricultural labourers in each group; the average of the *a* figures with the *c* weights is 12s. 5d., of the *b* figures with the *c* weights is 12s. 4d. Again, weight the figures for each county in column 4 with the numbers in column 5, the most obvious method of all; the average is then 12s. 4d. Again, take the simple average of the district averages *a* and *b*, that is, give each of the eight districts equal weights; the averages are 12s. 4¾d. and 12s. 3¼d. Or take the simple average of column 3, counting Yorkshire and Wales each as one county; it is 12s. 8d.

To obtain new groups, take as weights not the number of agricultural labourers, but the total population of the districts, the numbers marked *d*. Exclude the population of London as exerting a preponderating influence unconnected with agriculture. A new factor is now introduced, for population is greatest in the manufacturing districts, where agricultural labour is of compara-

Agricultural Wages in 1869-70. *To Illustrate Various Methods of Weighting, and their Results.*

	1. Michaelmas 1869.	2. Lady Day 1870.	3. Average of Cols. 1 and 2.	4. Average of Col. 2×2 and Col. 1.	5. No. of Agricultural Labourers in Counties. Unit 1,000.	6. Whole Population in Groups of Counties. Unit 100,000.
	s. d.	*s. d.*	*s. d.*	*s. d.*		
Sussex	12 3	12 0	12 1½	12 1	34	...
Surrey	14 0	13 6	13 9	13 8	16	...
Kent	14 6	14 0	14 3	14 2	44	...
Hants	11 0	10 6	10 9	10 8	32	...
Berks	12 0	10 0	11 0	10 8	22	...
Average	...	...	*a* 12 4½	*b* 12 3	*c* 148	*d* 22
Herts	14 7	11 10	13 2½	12 9	20	...
Northants	12 6	11 6	12 0	11 10	23	...
Hunts	16 0	11 0	13 6	12 8	9	...
Bedford	13 0	12 0	12 6	12 4	17	...
Camb.	11 0	12 0	11 6	11 8	24	...
Average	...	...	*a* 12 6	*b* 12 3	*c* 93	*d* 14
Essex	12 6	11 0	11 9	11 6	45	...
Suffolk	10 6	11 0	10 9	10 10	41	...
Norfolk	11 6	11 6	11 6	11 6	44	...
Average	...	...	*a* 11 4	*b* 11 3	*c* 130	*d* 12
Wilts	11 0	10 3	10 7½	10 6	26	...
Dorset	9 6	10 3	9 10½	10 0	17	...
Devon	10 0	10 3	10 1½	10 2	34	...
Cornwall	11 0	11 0	11 0	11 0	17	...
Somerset	11 0	10 6	10 9	10 8	31	...
Average	...	...	*a* 10 6	*b* 10 6	*c* 125	*d* 19

	1. Michaelmas 1869.	2. Lady Day 1870.	3. Average of Cols. 1 and 2.	4. Average of Col. 2×2 and Col. 1.	5. No. of Agricultural Labourers in Counties. Unit 1,000.	6. Whole Population in Groups of Counties. Unit 100,000.
	s. d.	*s. d.*	*s. d.*	*s. d.*		
Stafford	13 0	13 0	13 0	13 0	19	...
Gloucester	11 9	10 9	11 3	11 1	22	...
Hereford	10 3	10 0	10 1½	11 1	12	...
Salop	11 0	11 6	11 3	11 4	21	...
Worcester	13 6	11 0	12 3	11 6	15	...
Warwick	13 6	12 0	12 9	12 6	20	...
Average	...	...	*a* 11 9	*b* 11 7	*c* 109	*d* 27
Leicester	14 0	13 0	13 6	13 4	15	...
Rutland	12 6	12 0	12 3	12 2	3	...
Lincoln	14 0	13 6	13 9	13 8	49	...
Notts	13 6	13 0	13 3	13 2	16	...
Derby	13 6	14 0	13 9	13 10	8	...
Average	...	...	*a* 13 3½	*b* 13 3	*c* 91	*d* 14
Cheshire	13 6	13 6	13 6	13 6	18	...
Lancs.	15 0	15 0	15 0	15 0	30	...
Yorks, W.	19 0	15 3	17 1½	16 6	30	...
Yorks, N.	17 4	13 6	15 5	14 9½	16	...
Durham	16 6	16 0	16 3	16 2	8	...
Northumberland	19 6	16 6	18 0	17 6	12	...
Cumberland	15 0	15 0	15 0	15 0	10	...
Westmoreland	16 3	15 6	15 10½	15 9	3	...
Average	...	...	*a* 15 9	*b* 15 6	*c* 127	*d* 72
Monmouth	12 6	13 9	13 1½	13 4	6	...
Glamorgan	14 6	14 6	14 6	14 6	5	...
Caermarthen	12 4	11 6	11 11	11 9½	4	...
Pembroke	11 0	10 0	10 6	10 4	4	...
Cardigan	9 0	8 6	8 9	8 8	5	...
Brecknock	12 0	12 0	12 0	12 0	4	...
Radnor	10 0	10 0	10 0	10 0	2	...
Carnarvon	12 0	12 0	12 0	12 0	5	...
Average	...	...	*a* 11 7	*b* 11 7	*c* 35	*d* 14

tively little importance, but receives high wages; these high wages have undue weight, and the average of the figures *b* with weights *a* is brought up to 13s. 1¾d. The "median" of the county averages in column 4 is 12s. 1d. If column 4 is rewritten correct only to the nearest 1s., and column 5 to the nearest 10,000, the weighted average is 12s. 5d. If column 3 is weighted with random numbers quite unconnected with the problem, viz., the successive digits in the third decimal places of the logarithms of the numbers 2 to 46, the average is 12s. 10¾d. The reader may try any other system of logical or absurd weights, and he will find that unless there is some bias in the selection of weights, or great preponderance is given to a few counties, that the average will be little affected.

Since the true system of weights which would reduce the general average to our definition must be allied to some of those here adopted, and can hardly show greater divergence from 12s. 4d. than these do, we may feel confident that the true average is within, say, 3d. of this figure. The original items varied from 8s. 6d. to 19s.; the averages, even those based on the most extravagant methods, are contained by the limits 12s. and 13s. 1¾d. Without some such argument as this we should have no clue to the magnitude of the error introduced by erroneous weights. It is never safe, however, to assume that weights can be neglected, and an unweighted average used, without first examining the group in question, trying various systems, and seeing that the resulting average is stable. This will only be the case if there is no connection between the size of the quantities and the true magnitude of the weights. Thus if we are dealing with wages in towns, and are calculating the average for all towns taken together, we shall obtain too small a result if we ignore weights and count all towns as equal, for the higher wages are paid in the larger towns. Thus, as on pp. 134-5 below, the average of the recognised wages of 117 branches of the Amalgamated Society of Engineers was 32s. 4d. in 1891 if we count all the branches as equal; but was up to 33s. 4d. if we weight the wage at each of the branches with the number of members belonging to it. But, though we cannot neglect weights entirely in such cases, we need to make only a very rough estimation for them if there is no preponderating influence exerted by a small minority of places. In this case London, with a wage higher than any other district, except Dartford and Enfield Lock, and

Weights cannot in general be neglected, but only the errors in their estimation.

with nearly one-sixth of the total number of members dealt with, exerts such an influence. If, giving London its due importance, we take as weights the numbers belonging to the branches to the nearest hundred, we obtain the average 33s. 6d., practically the same as before. Each group for which an average is to be calculated must be treated on its merits; in many cases the weights may be neglected entirely; in nearly all cases, where the group consists of many items, even moderately large errors in computing weights may be neglected. Examination of the data will generally determine the importance of such errors.

This principle is of great importance. In many cases the true weights are incalculable or even undefinable; but now it is seen that, given certain conditions, there is no need to calculate or define the weights; in many other cases the weights cannot be known exactly, but exactness is not necessary. No system of weights, however, can remove an original bias common to all the items. If, for example, wages throughout were 1s. less than here reckoned, the calculated average would be 1s. too high. So we arrive at a very important precept: *in calculating averages give all care to making the items free from bias, and do not strain after exactness in weighting.*

C. THE MODE.—We pass to the consideration of two other means in common use among statisticians but unfortunately not yet consciously introduced into common parlance. There are, however, some popular phrases which, if they have any definite meaning, very nearly resemble the averages in question. When we hear of the average clerk, the average working-man, the phrases admit many interpretations. In some way these persons are supposed to be types of their kind. The average clerk may be supposed to mean the one who receives the average income of all clerks, whose expenditure on necessaries and on luxuries is the average of all of his class, who takes the average amount of interest in his work, is of average ability and average age. It will be seen that this clerk is ideal, and not to be found in any random assembly of half-a-dozen; for each of these will have some peculiarity, some quality in which he differs from the average; the average man of the newspapers does not exist in the flesh, but is an imaginary person to whom certain attributes are attached.

The average man.

Quetelet's average man is familiar;* he is of average height,

* See Quetelet's *Physique Sociale;* and Edgeworth in *Statistical Journal*, December 1893.

weight, strength, girth and lung capacity, with eyes of normal range and medium tint; but he is a more satisfactory model than the newspapers' average, for in regarding him we see the type from which all other men may be supposed to have deviated; the creature that would have been produced if all disturbing causes were removed. That any actual person should answer exactly to all these standards is of course in the highest degree improbable.

Quetelet's average man.

Quetelet refers neither to the arithmetic average, the median nor mode, but to a mean about which all the similar measurements are grouped in accordance with a definite law, the obedience of anthropometrical measurements to which was his chief theme.

The newspaper average, on the other hand, seems to be the mode, the position of the greatest density, which may be explained as follows:—Referring back to the table of American wages, p. 91, or the table on next page, it will be noticed that in looking down column 2 we find the numbers increase till we come to 685 (between $1.15 and $1.24), and then after fluctuations diminish. This number, 685, is the greatest which occurs in any 10-cent group; and its position is called the *mode*, or the position of *greatest density*, or the position of the *maximum ordinate*, or the rate is spoken of as *predominant*.

The mode.

In this column 2 we have, however, 14 maxima in the correct sense of the word, the numbers rise and fall with little regularity, and there are 14 modes of which that at $1.15-$1.24 is the most pronounced. But if the groups are made wider, and the numbers entered as in column 6 in half-dollar limits, there are only three modes, or if we neglect the small group of 8 at $5.00 only two. The position of the largest group of 1,472 is not at once assignable more closely than as between .75 and 1.25. We can get a little closer by the following method:—

Method of determining the mode.

Numbers earning as much as	$0.65,	and not as much as	$1.15	944
"	"	"	0.75	"	"	1.25	1,472
"	"	"	0.85	"	"	1.35	1,458
"	"	"	0.95	"	"	1.45	1,747
"	"	"	1.05	"	"	1.55	2,012
"	"	"	1.15	"	"	1.65	1,780
"	"	"	1.25	"	"	1.75	1,297
"	"	"	1.35	"	"	1.85	1,527
"	"	"	1.45	"	"	1.95	1,127
"	"	"	1.55	"	"	2.05	934

Determination of the Mode.

Numbers of Wage-Earners from the Senate Report, 1893, U.S.A.

			In 20-Cent Groups.		In 30-Cent Groups.		In 50-Cent Groups.	
From $.25	to	.34	1	16	74	75	159	317
" .35	"	.44	15					
" .45	"	.54	59	144	242	355	439	725
" .55	"	.64	85					
" .65	"	.74	157	270	282	674	1,190	1,472
" .75	"	.84	113					
" .85	"	.94	169	370	505	1,242	1,023	2,012
" .95	"	1.04	201					
" 1.05	"	1.14	304	989	784	740	603	1,297
" 1.15	"	1.24	685					
" 1.25	"	1.34	99	557	924	660	376	934
" 1.35	"	1.44	458					
" 1.45	"	1.54	466	538	274	343	399	640
" 1.55	"	1.64	72					
" 1.65	"	1.74	202	531	387	660	372	970
" 1.75	"	1.84	329					
" 1.85	"	1.94	58	331	418	310	146	506
" 1.95	"	2.04	273					
" 2.05	"	2.14	45	310	298	180	64	322
" 2.15	"	2.24	265					
" 2.25	"	2.34	33	134	297	181	226	198
" 2.35	"	2.44	101					
" 2.45	"	2.54	196	209	176	233	27	285
" 2.55	"	2.64	13					
" 2.65	"	2.74	163	165	17	32	242	254
" 2.75	"	2.84	2					
" 2.85	"	2.94	15	144	134	82	11	114
" 2.95	"	3.04	129					
" 3.05	"	3.14	5	52	59	85	93	96
" 3.15	"	3.24	47					
" 3.25	"	3.34	12	12	221	3	3	6
" 3.35	"	3.44	0					
" 3.45	"	3.54	221	226	21	4	4	9
" 3.55	"	3.64	5					
" 3.65	"	3.74	16	27	11	1	0	
" 3.75	"	3.84	11					
" 3.85	"	3.94	0	82	82	8	8	8
" 3.95	"	4.04	82					
" 4.05	"	4.14	0	3	3	8	8	
" 4.15	"	4.24	3					
" 4.25	"	4.34	0	0	3	1		
" 4.35	"	4.44	0					
" 4.45	"	4.54	3	4	1			
" 4.55	"	4.64	1					
" 4.65	"	4.74	0	0	0			
" 4.75	"	4.84	0					
" 4.85	"	4.94	0	8	8			
" 4.95	"	5.04	8					
" 5.05	"	5.14	0	0	8			
" 5.15	"	5.24	0					
" 5.25	"	5.34	1	1	1			
			0					

Now the greatest number is in the group \$1.05-\$1.55, and the "mode" may be stated as *near* the middle point of the group, viz., \$1.30, not *at* this point, for there are only 99 wage-earners in the group \$1.25-\$1.34.

Another method of approximating to the mode may be illustrated as follows:—When the numbers are tabulated in 10-cent groups, as on p. 91, the mode is quite indeterminate; in 20-cent groups the successive numbers beginning at .25-.44 are 16, 144, 270, 370, 989, 557, 538, 531, &c., and the number 989 (in the group \$1.05-\$1.24) is a distinct mode; if we begin the 20-cent groups at .35-.54, the numbers are 74, 242, 282, 505, 784, 924, 274, &c., and 924 (in the group \$1.35-\$1.54) is a mode; by this double tabulation it is seen that the 20-cent grouping does not decide the mode. In 30-cent groups we have 355, 674, 1,242 (\$1.15-\$1.44), 740, &c., if we begin with \$.55-\$.84; we have 439, 1,190 (\$.95-\$1.24), 1,023, &c., if we begin with \$.65-\$.94; and 483, 1,088 (\$1.05-\$1.34), 996, &c., if we begin with \$.75-\$1.04: the modes by each of these groupings lies in a group which contains \$1.15 to \$1.24, and this smaller group may be assumed to contain the mode, which is thus at or near \$1.20. The example here taken is drawn from a group of very irregular figures, which specially illustrate the difficulties. The method just adopted may be summarised thus:—Tabulate the figures again and again in gradually widening groups till regularity is obtained; then examine again the groups which have the selected width and see if the mode is shifted when the lower limit of the grouping is moved; if it is shifted the groups are not wide enough; if it is not, the mode is in the smallest group common to the larger equal groups which all contain it. A more accurate diagrammatic method is described on p. 154.

Indefiniteness of the position of the mode.

Even when our numbers are initially regular, it is seldom easy to determine the mode exactly. The difficulty is best seen by an example. Suppose that we have the following returns as to heights of a large number of men:—

67	in.	- -	455
67¼	„	- -	475
67½	„	- -	490
67¾	„	- -	500
68	„	- -	485
68¼	„	- -	467
68½	„	- -	445

At first sight the mode appears to be at $67\frac{3}{4}$ in. exactly; but it must be remembered that even in accurate measurements all heights within $\frac{1}{8}$ in. of $67\frac{3}{4}$ in. will be entered as $67\frac{3}{4}$ if the measurements are taken to the nearest quarter inch, or will have been tabulated in this way if the measurements were more accurate. Hence $67\frac{3}{4}$ in. in reality stands for from $67\frac{5}{8}$ to $67\frac{7}{8}$ in. If the 500 heights so entered were distributed uniformly through this interval, the mode might be given with $67\frac{3}{4}$ in. with fair accuracy; but there are signs in the figures that the mode is below this. Suppose that the figures in reality come from the following measurements:—

From $67\frac{1}{4}$ to $67\frac{3}{8}$ in.	238	} 483 at $67\frac{3}{8}$ in.	
„ $67\frac{3}{8}$ „ $67\frac{1}{2}$ „	245		
„ $67\frac{1}{2}$ „ $67\frac{5}{8}$ „	245	} 495 at $67\frac{5}{8}$ „	
„ $67\frac{5}{8}$ „ $67\frac{3}{4}$ „	250		
„ $67\frac{3}{4}$ „ $67\frac{7}{8}$ „	250	} 493 at $67\frac{7}{8}$ „	
„ $67\frac{7}{8}$ „ 68 „	243		
„ 68 „ $68\frac{1}{8}$ „	242		

and that these had been tabulated as in the last column, the mode would appear as $67\frac{5}{8}$ in.; while the same figures tabulated as before gave it as $67\frac{3}{4}$ in. The probability of some such shifting is seen from the original grouping, where the number at $67\frac{1}{2}$ in. is greater than that at 68 in. From this discussion we may see that the mode is always a little indefinite, depending on the width of the groups in which the items are tabulated, and on the exact position of the limits of the groups. As the items we deal with become more numerous, we shall find regularity when they are tabulated in narrower groups, and the mode can be assigned with greater accuracy. A more satisfactory method of determining the mode is that given on p. 155.

Now is the "average workman" the man who earns \$1.73 per diem, the simple average of the whole group on p. 120, or a man making \$1.20 the mode? In ordinary speech the latter is meant. The "average clerk" is not the one whose measurable qualities are an arithmetic mean of all similar qualities, but one whose qualities are found in the same degree in the greatest number of his fellows. There are more clerks who read the evening paper than who read Homer, more who go to music-halls than to oratorios, more whose

The "average man."

incomes are £100 than £500, more who live four miles from the City than one or twenty. Even with this explanation the average man is not a real creature, for fortunately no individual has no qualities out of the common. The fact that the average is a pure abstraction is of importance directly we apply statistics to actual affairs; these American workpeople cannot be legislated for in the mass as if they all earned $1.20, or as if those who were alike in this did not differ in other respects, even doing very varying quantities of work for this wage. No single measurement expresses completely even the economic condition of a group of workmen, but if we are taking a single measurement, that of the "mode" is often the most useful. It is at the mode that we find the greatest number of whose greatest good we may be thinking. Whereas the arithmetic mean and the "median" may correspond to no reality but be merely numerical conceptions, the mode is precisely that number for which most instances can be found. It shows the commonest result, that most often obtained, and is of very general application. For an intending passenger by train or 'bus, it is more important to know the most ordinary than to know the average number in a compartment. The mode rather than the average in chest measurements is the number most suitable for the ready-made clothier. For providing a post-office or a store, the mode in postal orders or prices of tea needs to be known rather than any other average. Even the favourite coin in a collection may show the spirit of the congregation better than the arithmetic average of their contributions. In these last instances it may be noticed that the mode is quite definite.

Importance of the mode.

A special feature of the mode is that it is entirely uninfluenced by extremes. A cheque for £1,000 in a collection disturbs the arithmetic average, but not the mode. The incomes of a small number of millionaires and an army of paupers may have the same arithmetic average as a nation composed entirely of people moderately well off; but the modes will be very different in the two cases. In considering the change year by year in a group of figures, as for instance, the wages of a large group of workmen, we cannot tell, if we take the arithmetic average as our criterion, whether an improvement is due to a levelling up of the badly paid or a rapid increase for those who were already well off, while the mode will show the changing position of the main body. Mr Booth's "London" is crowded

Advantages of the mode.

with instances of the use of the mode. Each age diagram shows the mode in ages for an occupation; each wage list that in wages. His whole description of Class E, the typical workmen of modern towns, is based on the same principle. His measurement of social status, based on the number of rooms occupied or servants employed, can be used easily for stating the mode (four rooms to a family and no servant) but not any other average.

Shortcomings of the mode.

An objection to this average is that there are many groups of figures to which it is not applicable. If we have a very irregular group of numbers with no particular type, such as the populations of towns in England, the mode would be quite indefinite, and would give no information of importance. The use of the mode is to indicate the type from which other figures may be regarded as diverging. Thus, in these wage figures, the type is about $1.20, and other examples lie on either side, wages of men who have for some reason or other more or less than the normal degree of skill or opportunity. If there is a type, as in Quetelet's instances, the mode will show it. The mode only tells us one fact, however, about each type, whereas the methods already given (p. 92) show us several.

D. THE MEDIAN.—When we are dealing with a group of persons or things, each of which possesses some measurable attribute, such as height or wage, we can choose certain quantities which describe the group in brief. Suppose all the items arranged in a series in ascending order of the magnitude of this attribute; the magnitude appertaining to the item half way up the series is called the *median*.*. Thus if in a group of wage-earners 200 earn less than 20s. 3d., one earns 20s. 3d., and 200 more, 20s. 3d. is the median wage. There are as many items below 20s. 3d. in the supposed series as above it. The magnitudes one-quarter and three-quarters up the series are called the *quartiles;** those one, two . . . nine-tenths up are the *deciles;* those one, two . . . ninety-nine hundredths up are the *percentiles*. The median is more definite in position than the *mode*. When we are dealing with exact measurements, if we have an odd number of items it is the middle one, if an even number, it lies between the two middle items, which are in general near together, or coincides with them if they are equal. If the magnitudes are not given exactly, but as within small limits, we can by the method described on pp. 127-8 make a good estimate of the actual values of these averages. The median is not affected by

* These quantities have already been used in tabulation, p. 95.

exceptional entries at all; the existence of any number of millionaires has no more effect on the median income than of an equal number of any other persons whose incomes are above the median. For many purposes it is of course necessary to allow these extreme instances more weight than those which are nearer the average; but the arithmetic average often gives them undue weight for this democratic age, since a single millionaire can counterbalance thousands of ordinary working men. A further advantage is that it is extremely simple to find, not needing much arithmetical work, for we need not do more than count those well above and well below the average, and look more carefully at those near it.

No need for complete information.

There is a yet more important advantage in the use of the median; it can often be found exactly, when our information as to the items in question is neither accurate nor complete. This will be clear from one or two examples. It may be that in the "wage census" 100,000 persons, whose wages were far below the average, did not come into the returns at all, and it is very difficult to estimate their effect on the arithmetic average, for want of information as to their earnings; but to find the median exactly, we need only know their number, not their earnings; and if we can only assign a maximum for their number, we still can place the median within narrow limits. The addition of 100,000 men with wages below 15s. to the general summary for the 356,000 men, would still leave the median in the group 20s. to 25s. where it already is; the change would be very marked, however, in the lower deciles and quartiles, and the arithmetic average would be lowered by at least 2s. 1d. The same argument applies to incomes; information is often very deficient, but it is in many cases possible to assert that a number of men, whose exact income is unknown, receive above a certain assigned sum, or even between two assigned limits, which is all we need to know about them to determine the median, if it lies below the lower limit.

Again, in tracing the history of wages throughout the century it is often very difficult to find the correct average, but at the same time it is frequently possible to say that a very large class of men earned below, say, 15s. a week, and another very large class above 30s. whose wages we do not exactly know, and a more definite number between 15s. and 20s., and 25s. and 30s.,

and in order to find the median all we need to do is to investigate more exactly the wages between 20s. and 25s., and even if we have not complete information here, we can still say that the median certainly lies between certain narrow limits. There is yet another advantage, perhaps more important, that the median is applicable to quantities which are not capable of measurement at all. This development is especially due to Mr F. Galton.* Suppose it to be required, for example, to find among a large class of boys the average in intelligence. It is clear that it is not easy to find the arithmetic average of a quantity which cannot be properly measured even by the most elaborate system of marks, but on the other hand it would not be at all difficult with a class of, say, twenty boys, to place them in order of intelligence without committing oneself to such a statement as that A.'s cleverness was 25 per cent. more than B.'s; and the tenth or eleventh boy in this arrangement would show the style of boys in the class, at least as well as any other average. The disadvantage of this method, the reason why it is not universally applicable, is that the median of a series of observations may be totally removed from its type, and in fact may not be situated near any of the different objects which are observed. Thus, if we had two large groups of wages of a thousand men between 15s. and 25s., and another thousand between 35s. and 45s., the median would give us any position between 25s. and 35s., where as a matter of fact not a single wage-earner would be found. The median is then chiefly useful when we are dealing with a series of objects of which the main part lie fairly close together; a few extremes do not affect it.†

Incommensurable quantities.

Disadvantages.

If m is the median and a the arithmetic average of n quantities $x_1, x_2 \ldots x_n$, and we call $x_1-z, x_2-z \ldots$ the *deviations* of the x's from any quantity z, then m is the value of z which makes the sum of the deviations (all taken positively) a minimum, a is the value which makes the sum of the squares a minimum. The first statement is easily verified; for the second, we notice that $\Sigma x = na$, and that $\Sigma(x-z)^2 = \Sigma x^2 - na^2 + n(a-z)^2$, which is a minimum when $z=a$.

The following table shows the description of 76 items by the help of the various averages now described:—

* See, for instance, *Natural Inheritance*, p. 47.

† On the relative advantages of this, and a more mathematical method, see *Yule* and *Galton* in the *Statistical Journal* for 1896, especially pp. 392-398.

MEASUREMENTS OF BOYS OF AGES 13 TO 15 YEARS.

No.	Age.	Height.	Weight.	No.	Age.	Height.	Weight.	Tabulation of Weights.
	yrs. mth.	ft. in.	t. lb.		yrs. mth.	ft. in.	st. lb.	
1	14.1	4.11½	6.0¾	39	14.7	4.11½	6.3½	Arithmetic average, 6 st. 1½ lbs.
2	14.9	4.10	5.7	40	13.1	4.11½	5.7	
3	14.7	5.5¼	7.5	41	14.3	4.11	6.4¼	
4	13.11	5.0	6.3¾	42	13.3	4.4½	4.11½	The same, when weights are entered only to nearest stone, 6 st. 1⅛ lbs.
5	14.11	5.3¾	8.0½	43	14.3	5.3	6.7½	
6	14.7	4.10	5.0	44	13.6	5.1½	6.13¼	
7	14.3	4.10	6.7	45	14.2	4.8¾	6.0¾	
8	14.9	5.5	8.5¼	46	13.5	5.2	7.4	
9	14.11	4.9½	5.12¼	47	13.8	5.2½	6.11	
10	14.3	4.11¾	6.11½	48	14.6	5.4	7.4½	Median, 6 stones 1⅛ lbs.
11	13.4	4.7	5.1½	49	14.8	5.1½	6.10	
12	14.7	5.3¾	7.8¾	50	13.3	4.8¼	5.0	
13	13.8	4.7¾	5.3	51	13.0	5.1¼	6.7	Quartiles, 6 st. 9⅞ lbs., 5 st. 6⅛ lbs.
14	14.5	5.2½	7.8¾	52	13.10	4.11½	7.3¼	
15	14.4	5.0	6.0	53	14.8	4.11½	6.9¾	
16	13.6	4.9	5.6	54	13.8	4.5¾	4.9½	Average of quartiles, 6 st. 1 lb.
17	14.0	5.2½	7.7½	55	14.8	5.4½	7.0	
18	13.0	4.8¼	5.3	56	14.0	4.10	6.2½	
19	14.7	4.11	6.12¾	57	13.10	4.9	5.5	Half of the examples lie within 9 lbs. of median.
20	14.10	5.1	6.9	58	13.2	5.0⅝	6.4	
21	13.9	4.11	5.11	59	13.6	4.7	5.2½	
22	14.10	4.8¾	5.11	60	13.0	4.9	5.9¾	
23	13.4	4.9½	5.8¾	61	13.3	4.8¾	5.5¼	Mode is between 6 st. and 6¼ st.
24	13.1	5.2¼	6.1	62	13.5	4.8½	6.5¾	
25	14.0	4.6¼	5.6½	63	13.10	5.5½	7.10½	
26	14.6	5.3½	7.6½	64	13.1	4.8¾	6.2¼	Average weight between ages 13 and 13½ years, 5 st. 9½ lbs.; 13½ and 14 years, 5 st 13⅛ lbs.; 14 and 14½ years, 6 st. 3⅜ lbs.; 14½ and 15 years, 6 st. 8⅔ lbs.
27	14.3	5.0¼	5.11¾	65	13.10	5.4	7.2	
28	13.9	4.9	5.11	66	14.0	4.9	5.0½	
29	13.4	5.1¼	5.9	67	13.3	4.7	5.0	
30	14.4	5.1	6.8½	68	13.8	4.11	6.1¼	
31	14.10	4.9½	4.7¼	69	13.7	4.11¾	6.4¼	
32	13.2	4.9½	5.13¼	70	13.11	4.8	4.4½	
33	14.1	4.8¾	5.8½	71	13.11	4.8	4.4½	
34	13.10	5.2¼	6.8¼	72	13.2	4.7¾	4.10	
35	14.0	4.11½	5.7	73	14.0	4.11	6.5	
36	14.4	4.11	6.5	74	13.3	4.3½	4.1¼	Heights may be tabulated in the same way.
37	14.8	4.11	6.0¾	75	13.3	5.0	7.2¾	
38	13.7	5.0¾	6.2	76	13.7	4.8¼	5.6	

Graphic method. A graphic method of finding the median closely is given by Mr Galton in the *Report of the Anthropometric Committee* of the British Association, 1881, p. 247; and is illustrated by the diagram facing the next page.

On a horizontal line mark off equal intervals representing units of measurement, say inches. On a vertical scale mark off equal intervals representing the number of instances, *e.g.*, persons whose heights are measured. Beginning at the lowest, say 51¼ inches, on an imaginary vertical line mark as many dots at equal intervals on the vertical scale as there are

persons at that height, so that each dot represents one person. From the highest dot thus marked, suppose a horizontal line drawn till it is over the next height division, $51\frac{1}{2}$ inches, and with this new base proceed as before, marking each instance at $51\frac{1}{2}$ inches by a dot vertically above the $51\frac{1}{2}$-inch mark. Next draw a connected line through the middle points of the consecutive vertical rows of dots; if there is an odd number of dots, the middle one is taken as the middle point; if an even number, the middle point is half-way between the middle ones.

On the vertical scale mark the positions of the median, quartiles, &c., obtained by dividing the distance representing the total number of instances into appropriate parts, and through these points draw horizontal lines to intersect the connected line already drawn. The points of intersection lie vertically above the heights required, as marked on the horizontal scale.

Now it may be assumed that the heights of all persons returned at, say, $58\frac{3}{4}$ inches, are in reality evenly distributed between the limits $58\frac{5}{8}$ and $58\frac{7}{8}$ inches, heights lying within which would be so returned; and it can be verified that the construction just given shows the place of the median, deciles, &c., almost exactly on this hypothesis.

E. GEOMETRIC MEAN.—It is not necessary to give a long discussion of the geometric or logarithmic mean, for its application is limited to a small class of figures which will be best dealt with at a later stage.* It was used by Jevons in his essay on the *Fall in the Value of Gold*, but he did not justify or explain its use.† If we have n quantities $a_1, a_2, \ldots a_n$, their geometric mean is $\sqrt[n]{a_1 . a_2 . \ldots a_n}$ Its chief advantage is that the influence of large numbers is diminished and of small numbers increased, when the geometric mean is employed instead of the simple average. Suppose all the following groups of numbers represent price levels of various commodities as percentages of their height at a previous date:—

Numbers.	Arithmetic Mean.	Geometric Mean.
80, 160	120	113
80, 80, 100, 324	146	120
20, 20, 80, 80, 120, 120	73·3	57
20, 20, 80, 80, 100, 100, 120, 160, 324, 972	198	104

* See p. 223, *infra.* † But see *Statistical Journal*, 1865.

A consideration of the last list leads to the conclusion that the general rise of price cannot be 98 per cent., while 4 per cent. may reasonably represent it. A tentative rule may be suggested: when the geometric mean differs much from the arithmetic it should be preferred. It should be calculated with the help of logarithms.

F. STATISTICAL COEFFICIENTS. — Before leaving the subject of averages, we must pay some attention to "statistical coefficients." A statistical coefficient is a number, whole or fractional, by which a total (*e.g.*, population) must be multiplied to give an allied number (*e.g.*, number of births). Thus, if the birth-rate is 40 per 1,000, the coefficient is .04. These coefficients play an important part in ordinary statistics and a very interesting *rôle* in the application of the law of error to demography. The population may increase or diminish, but the coefficients relating to certain numbers remain almost unchanged,* and by their use the statistics of different countries may be compared, and numbers for future years can be forecasted in some cases with marvellous accuracy; the numbers of births, marriages, deaths in 1901 can be written down before their occurrence as exactly as they are needed, subject only to the chance of some great catastrophe. Coefficients can be formed for births (in various districts), for deaths (according to age, profession, or disease), for marriages (at various ages), for suicides, crimes, accidents, consumption of various commodities; if the preliminary data could be obtained, for the number of persons crossing Westminster Bridge in the year, the number of visitors to the Monument, the number of umbrellas left in the train, and so on; the list could be prolonged indefinitely. The more important coefficients are calculated for most civilised countries, and the rates on which they are based published in statistical abstracts. A knowledge of them is necessary for statistical investigations.

A useful caution is given by Dr Bertillon.† In order that a coefficient may obey the laws of coefficients closely, the number to which it is to be applied should not be that of the total population, but the number of persons or things capable of affording an instance of the resulting total. Suppose m to be this number of persons, c the coefficient, n the resulting total, then $n = cm$ and $c = \frac{n}{m}$. Thus,

Calculation of coefficients.

* *See infra*, Part II. † *Cours élémentaire*, p. 94 *seq.*

for the marriage rate, m should not be the total population, but the number of marriageable people. The importance of this rule is, however, not great as far as simple calculations are concerned; for the less accurate coefficient can be easily seen to be nearly constant. Suppose M to be the total population, m the number of marriageable people, n the number of marriages. Let c_1 be the coefficient for calculating the number of marriageable people, then $c_1 = \frac{m}{M}$; c_2 the coefficient for marriages on Bertillon's principle, then $c_2 = \frac{n}{m}$. Let $c_3 = \frac{n}{M}$, the more usual coefficient for marriages. $c_3 = \frac{n}{m} \times \frac{m}{M} = c_1 \times c_2$. Now if c_1 and c_2 are invariable, so also is c_3. If, however, one of the factors, say c_1, is more variable than the other, then c_3 varies as much as c_1, and the greater constancy of c_2 is not discovered, if only c_3 is calculated.

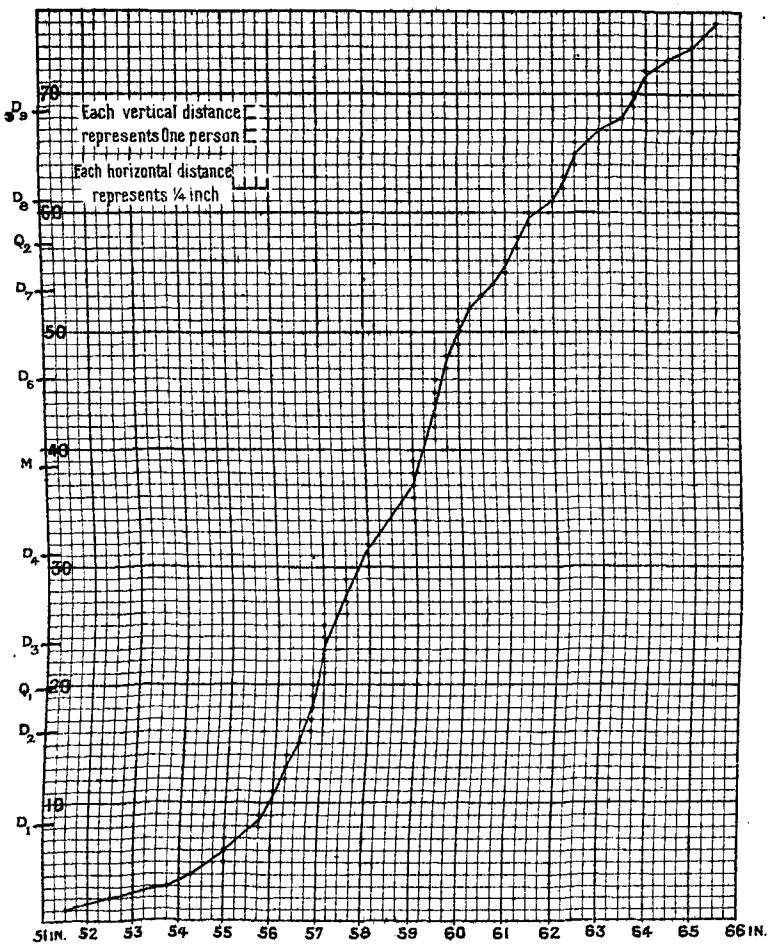
G. GENERAL.—The function of averages will now be clear; it is to express a complex group by a few simple numbers. The mind cannot grasp the magnitudes of millions of items at once; they must be grouped, simplified, averaged. The averages chosen must be those which will give the striking features and the essential characteristics of the group. Different methods will apply to groups of various classes; each must be taken on its own merits. A good and suitable average has the following characteristics:—*If there is a type it shows it; it gives due influence to extreme cases; it is not easily affected by errors or much displaced by slight alterations in systems of calculation; and it is easily calculated.*

The function of averages.

The relative positions of the different kinds of averages dealt with gives some information as to the general nature of the group to which they refer. The arithmetic average, median and mode, are close together, if the group is symmetrical. The arithmetic average is probably above the median, if we have a small group at a high degree. The arithmetic average is generally below the median, if there is an absence of high numbers, and a concentration a little above the average. The mode will be badly defined, if our group is not homogeneous. The mode will probably be below the arithmetic average, if there is a small group at a high degree. The mode is well marked, if the distribution is uniform. These rules are only tentative and easily nullified by exceptional circumstances.

GRAPHIC METHOD OF FINDING MEDIAN, QUARTILES AND DECILES (after Galton: Athropometric Committee: Brit. Assn.).

For the Heights of the 76 boys, between ages of 13 and 15, stated on p. 127.



Median $59\frac{1}{8}$ inches.

Quartiles $\begin{matrix}56.8.\\61.2.\end{matrix}$

'Probable error' 2.2.*

Deciles 55·5, 56.6, 57, 57.9, 63.6, 62, 60.6, 59.7.

Arithmetic average, 59.095.

Greatest density 57 or 59.

" " in smoothed curve would be about 58.

Geometric average 58.98.

CHAPTER VI.

SOME EXAMPLES OF THE USE OF AVERAGES IN TABULATION.

CHAPTER VI.

SOME EXAMPLES OF THE USE OF AVERAGES IN TABULATION.

IF our analysis of the nature and use of averages is complete, and if averages are of widely extended use, we should now be able to express almost any group of figures by a few well-chosen numbers of definite significance.

Application of averages

To apply a somewhat severe test at first, let us choose a familiar example from ordinary life, and consider how a suburban business man might test the merits of two railway systems, by one of which he intended to take a season ticket.

to train service.

The following table gives the train service between Leatherhead and London in 1898:—

TRAIN SERVICE—LEATHERHEAD TO LONDON.

Number of Minutes to Journey.

WATERLOO—

Down—60, 50, 52, 48, 47, 61, 50, 44, 48, 53, 45, 42, 45, 49, 43, 48, 42, 43.
Sundays—50, 50, 47, 49, 50.

Up—51, 46, 51, 48, 43, 44, 48, 48, 64, 45, 48, 47, 45, 47, 46, 47.
Sundays—48, 48, 51, 51, 51.

LONDON BRIDGE—

Down—67, 65, 65, 61, 74, 51, 56, 66, 65, 53, 59, 41, 49, 44, 58, 57, 56, 67, 80.
Sundays—67, 52, 66, 68, 88, 65, 65, 68, 65.

Up—69, 57, 53, 58, 54, 41, 58, 52, 42, 40, 55, 67, 79, 98, 69, 66, 68, 64, 71.
Sundays—72, 71, 69, 70, 62, 81, 73, 73.

VICTORIA—

Down—77, 65, 55, 76, 77, 88, 48, 53, 46, 69, 89, 54, 82, 71, 90.
Sundays—92, 45, 81, 84, 78, 61, 85, 83, 85.

Up—87, 65, 69, 69, 47, 48, 51, 83, 101, 58, 62, 61, 76, 103.
Sundays—81, 76, 80, 85, 85, 82, 94.

The following table gives us the necessary information:—

	London Bridge.	Victoria.	Waterloo.
	Min.	Min.	Min.
Average of four quickest trains	41	46½	42½
Lower decile	47½	48	43
Median	65	77	48
Mode	65	...	48
Number of trains on week days	38	29	34
General average	63	73	48

It is to be noticed that the statistical method is generally limited to one aspect of a problem; the question of punctuality might, indeed, be easily treated statistically, but the questions of comfort and relative picturesqueness of route will elude our analysis.

The next example shows a method of throwing into relief the characteristics of a typical group of sociological data.

Tabulation of wages returns.

The adjoining table gives the wages recognised by the Amalgamated Society of Engineers in many of their branches in 1862 and 1891.

AMALGAMATED SOCIETY OF ENGINEERS.—WAGES in 1862 and 1891, Weekly, exclusive of Overtime.

	1862. s.	1862. d.	1891. s.	1891. d.
Accrington	27	0	31	0
Ashford	33	6	30	0
Ashton-under-Lyne	29	3	34	0
Bacup	26	1	28	0
Barrow-in-Furness	31	0	34	9
Bath	29	0	31	0
Bedford	27	0	29	0
Bilston	28	0	30	0
Bingley	24	0	29	0
Birkenhead	29	0	35	6
Birmingham	32	0	36	0
Blackburn	27	6	32	0
Bolton	27	6	28 / 32	0 / 0
Bridgwater	24	6	24	0
Brighton	24	8¼	29	0
Bristol	31	0	32	0
Burnley	27	0	30	0
Burton-on-Trent	25	0	30	0
Bury	28	3	30 / 32	0 / 0
Cardiff	31	0	34	0
Carlisle	24	6	30	0
Chepstow	30	0	34	0
Chester	30	0	32	0
Chowbent	26	0	32	0
Colne	25	0	31	0
Congleton	24	0	28	0
Coventry	28	0	34	0
Crewe	29	4	30	0
Darlington	25	0	31	6
Dartford	34	0	38	0
Darwen	27	0	32	0
Derby	26	0	29	0
Doncaster	28	6	31	6
Dover	35	6	36	0
Enfield Lock	36	0	40	6
Exeter	23	0	28 / 32	0 / 0
Faversham	34	0	33	0
Folkestone	34	0	32	0
Frome	24	0	27 / 30	0 / 0
Gainsborough	27	6	28	0
Glossop	27	2	32	0
Gloucester	28	0	32	0
Grantham	28	6	30	4
Grimsby	28	0	32	0
Halifax	23	1	31	0
Hanley	28	3	32	0
Hartlepool	26	0	34	10
Heywood	27	0	30 / 34	0 / 0
Holyhead	32	0	28	0
Huddersfield	26	0	26	0
Hull	27	6	34	0
Hyde	30 / 28	0 / 0	30 / 28	0 / 0
Ipswich	28	6	28	0
Keighley	23	0	27	0
Kidderminster	28	0	30	0
Lancaster	25	0	32	0
Leeds	25	0	30	0
Leicester	26	0	31	6
Leigh	27	9	31	6
Lincoln	26	7	28	6
Liverpool	29	0	34	0
Llanelly	22	0	26	0
Macclesfield	24	0	29	6
Manchester	29	9	35	0
Mexborough	27	0	32	0
Middlesborough	25	0	34	0
Middleton	29	5	33	0
Milton and Elsecar	28	0	34	0
Neath	32	0	30	0
Newark	25	0	29	0
Newcastle	25	0	35 / 37	0 / 0

	1862. s.	1862. d.	1891. s.	1891. d.
New Holland	30	8	34	0
Newport	30	0	32	0
New Town (Stockport)	29	0	32	0
Newton Abbott	33	0	33	0
Northampton	26	0	32	0
Northfleet	36	0	36	0
North and So. Shields	26	0	35	0
Norwich	32	0	29	0
Nottingham	27	5	34	0
Oldbury	28	0	34	0
Oldham	29	0	33	0
Peterborough	28	6	33	0
Plymouth	32	0	33	0
Pontypridd	24	0	30	0
Portsmouth	35	0	34	0
Preston	27	0	32	0
Radcliffe Bridge	27	0	30	0
			32	0
Reading	28	0	32	0
			34	0
Ripley	26	0	26	6
Rotherham	27	6	32	0
Rugby	32	0	28	0
			32	0
Rugeley	24	11	30	0
St Helens	28	0	34	0
			36	0
Sheffield	28	0	36	0
Shipley	25	9	28	0
			30	0
Shrewsbury	30	6	32	0
Smethwick	28	0	35	0
Southampton	32	0	34	6
Sowerby Bridge	24	6	30	0

	1862. s.	1862. d.	1891. s.	1891. d.
Stafford	34	0	30	0
Stalybridge	28	3	32	0
Stockport	28	0	2	0
			34	0
Stockton-on-Tees	24	0	36	0
Stoke-on-Trent	29	0	32	0
Stroud and Thrupp	26	0	30	0
Swindon	31	6	31	6
Todmorden	26	0	28	0
Wakefield	25	0	30	0
Warrington	28	0	34	0
Watford	35	0	36	0
Wednesbury	26	0	31	0
Whitehaven	25	0	28	0
			36	0
Wigan	28	0	34	0
Wolverhampton	28	0	33	0
Wolverton	29	2	29	0
Worcester	31	0	30	0
Bermondsey	35	4		
Blackwall	34	0		
Bow	36	0		
Greenwich	34	0		
King's Cross	36	0		
Lambeth	35	8		
London, E.	35	0		
,, N.	35	10	38	0
,, S.	35	0		
,, W.	35	6		
Marylebone	33	0		
Stratford	35	0		
	33	6		
Tower Hamlets	36	6		
Woolwich	36	0		

The following figures show the same in brief:—

	1. 1862.* s.	1. 1862.* d.	2. 1891.* s.	2. 1891.* d.	3. 1891.† s.	3. 1891.† d.
Maximum	36	6	40	6	...	
Upper decile	35	0	38	0	38	0
Upper quartile	31	4	34	0	36	0
Median	28	0	32	0	34	3
Arithmetic average	28	10	32	4	33	4
Modes	28	0	30	0	...	
			32	0		
Lower quartile	26	0	30	0	31	6
Lower decile	24	6	28	6	30	0
Minimum	22	0	24	0	...	

* Each branch counting as 1.

† The numbers of members in each branch counted as receiving the wage recognised there.

If the rates at each branch were not those actually paid to all members, but their average while the actual wages were confined within small limits of that average, the figures in the last column would be little affected.

On comparing columns 1 and 2 it will be seen that not only have all the averages increased, but that since the lower decile and quartile have increased more rapidly than the upper, the lower half has also gained on the upper. Again the wages are grouped more closely in column 2 than in column 1.

We need a simple and general method of describing the distances of members of a group from their average and each other, in order to compare one group with another.

Measure of dispersion.

The fraction $\frac{Q_2 - Q_1}{Q_2 + Q_1}$, where Q_1 and Q_2 are the quartiles, supplies such a measurement, and may be called the *dispersion*. It is easily verified that this fraction increases when the ratio of the upper to the lower quartile increases, and that it is sensitive to changes in a group, and that it lies between 0 and 1.

For the above figures this quantity is .093 in 1862, and $\frac{34s. - 30s.}{34s. + 30s.} = .062$ in 1891. The dispersion diminished in that period.

Other suitable measurements of dispersion would be (in the notation of p. 126) $\frac{\Sigma(x-m)}{nm}$, all the terms in the numerator being taken positively, or $\sqrt{\left\{\frac{\Sigma(x-a)^2}{n}\right\}} \div a$. For the above figures in 1891 these become .078 and .097 respectively. The methods will in general give different results, often in the proportion here found, and the special method adopted must be adhered to during an investigation.

GROUP C OF TABULATION.—It was necessary to postpone the tabulation of non-numerical or descriptive answers till we had finished our discussion of averages. The following detailed example shows how the median, &c., can be used to give a short description of a large group of adjectival answers.

Tabulation of descriptive answers.

In 1891 the Amalgamated Society of Engineers obtained from all their branches answers to the question: To what extent is overtime worked? The branch secretaries sent answers which may be tabulated as follows:—

Answers.	Number of Branches.	Number of Members.
None	4	140
Not worked	1	78
Very little	23	4,836
To very limited extent	1	63
Very occasionally	1	350
A little on repairs	1	500
Little	2	73

Answers.	Number of Branches.	Number of Members.
2 hours when necessary	1	80
Seldom	1	59
Small extent	1	16
Seldom except on repairs	1	66
Only on repairs	2	216
Not much	6	1,125
On repairs	1	500
Not to any extent	3	644
Not to a great extent	2	162
Not general	1	7
Not systematically	2	43
In cases of breakdown or emergency	7	606
2 hours regularly	1	136
Chiefly on repairs	1	20
Occasionally	2	90
When necessary	1	348
Casually (*sic*)	2	142
A good deal on repairs	1	23
Maximum 18 hours in 4 weeks	1	1,000
Moderately	3	262
Systematically in good trade	1	200
Average about 5 hours a week	1	96
Considerably in marine shops	1	400
Systematically in dockyard	1	650
General	2	146
Systematically	1	693
Great amount	1	263
To a great extent	1	72
Excessively	1	550
9 hours a week	1	39
10 „	1	106
12 „ (maximum)	1	700
14 „ (when busy)	1	106
10 to 18 hours a week	1	5,000
Total	88	20,666
UNCLASSED :—		
No answers	36	5,114
As little as possible	1	250
Not so much lately	1	166
In machine shops for six months	1	60
In steel works	1	348

An inspection of the table here given will show sufficiently the method of tabulation. The position of most of the answers in an imaginary scale is fairly definite, except that it is not always obvious where the numerical answers should be placed; this must be decided either by internal evidence or practical knowledge of the trade. The same adjectives did not of course convey exactly the same numerical meaning to all the branch secretaries who used them, but it will be admitted that this tabulation gives a fairly clear view of the case, and that the method of medians and quartiles may be appropriately applied. Taking the member of a branch as the unit and neglecting the unclassed answers, the median is "Maximum 18 hours in 4 weeks" or "moderately," the lower quartile "Very little," and the upper quartile "14 hours when busy." Taking the branch as unit, the median is "Not much," the quartiles are "Very little" and "When necessary" or "Occasionally."

Explanation of table.

This method, which, with varying degrees of precision, is widely applicable, seems to afford the only way of comparing two such groups of answers. The precision attainable is to be measured by the distance through which the median can be shifted by making reasonable variations in the scheme of tabulation.

SUMMARISATION.—Now that we have the method of averages at our disposal we may use it for tabulating and summarising a group of figures.

Consider, for example, the answers to the questions issued by the Commissioners on Trade Depression in 1886.

Four of the questions were:—

1. Number of men in Society.
2. Number out of work in 1885.
3. Weekly wage in 1885.
4. Change in wages between 1865 and 1885.

The following table shows the answers given by the branch secretaries of the Amalgamated Society of Engineers:—

1. District.	2. No. in District, 1885.	3. No. Out of Work, 1885.	4. Current Wages, 1885.	5. Change between 1865 and 1885.
Belfast	1,100	130	28/ to 36/	Slight increase.
Coventry	2,500	230	31/6	Contract work—50 °/₀ decrease.
Dukinfield	170+	20+	31/	Slight increase.
Dundee	1,400	45°/₀	25/ skilled. 15/ unskilled.	Time work—1865, 22/; '72, 24/; '80, 26/; '83, 24/; '85, 25/.
Glasgow	28,000	4,000	26/	Time wages, 5 °/₀ above 1864.
Glasgow (St Rollox)	1,600	250	...	Rise in 1872-73 of 15 °/₀; 1885 same as 1865.
Hartlepool	1,200	400	31/6	Advance of 3/.
Glossop	135	10	32/	.....
Liverpool	280	38	...	Rise in 1872-73 of 7½ °/₀; 1885 same as 1865.
Monifieth	114	18	21/	Skilled work—1865, 24/; '76, 27/; '78, 25/; '83, 28/; '85, 25/.
Nottingham	4,000	600	34/minimum.	1865, 28/; 1885, 34/.
Oldham	1,600	96	33/ average.	Increase of 5 °/₀.
Oxford	45	...	33/	.....
Paisley	800	...	28/6	1865, 26/; 1885, 28/6.
Preston	630	40	28/	None.
Preston	900	120	28/	None.
Shipley	201	15	28/6 24/ non-unionists.	1865, 28/6; 1869-73, 32/; 1885, 28/6.
Sowerby Bridge	1,120	43	28/	1865-75, 25/6; 1875-85, 28/.
Sunderland	3,200	400	33/	1864, 27/; '74, 34/; 1875-85, between 31/ and 37/.
Swindon	6,050	2	31/6	.....
Ulverston	45	...	31/	1865, 26/; 1875, 31/.
Wednesbury	400	30	30/	Increase of 2/.
Workington	170	70	28 to 36/	Increase of 30 °/₀.

It is suggested that the following are the summary tables which should be inserted in a report dealing with the answers.

The figures are given here for only one society, but the tabulations are framed so as to include all.

TABLE I.—STATE OF EMPLOYMENT.

Name of Society.	Total Number * in Branches making Returns on Employment.	Number Out of Work.	Percentage Out of Work.	Median of the Percentages Out of Work in the Various Branches.
A.S.E.	55,170	7,142	13	12
O.S.B.				
&c.				

* Details of some of the most important branches should be added.

TABLE II.—CURRENT WAGES.

Name of Society.	Average of Wages in Branches.		Quartiles of Branch Wages.		Measure of Dispersion (*v.* p. 136).
	Unweighted.	Weighted.			
	s. *d.*	*s.* *d.*	*s.* *d.*	*s.* *d.*	
A.S.E. -	30 0	29 7	28 0	32 0	$\frac{1}{15}$
O.S.B. -					
&c. -					

TABLE III.

A. CHANGE OF WAGE BETWEEN 1865 AND 1885.

Name of Society.	Number of Branches showing				Median of Percentage Increases.	Percentages of Members in Branches showing			
	No Answer.	Decrease.	No Change.	Increase.		No Answer.	Decrease.	No Change.	Increase.
A.S.E.	4	1	5	13	10	11	4	6	79
O.S.B.									
&c.									

Verbal Summary.—In the great majority of cases a considerable increase of wage took place between 1865 and 1885, equivalent on the whole to a rise of about 10 per cent. The figures are not sufficiently definite to give an exact average.

TABLE III.—*B.* CHANGE OF WAGE BETWEEN 1865 AND THE MAXIMUM ABOUT 1873.

TABLE III.—*C.* CHANGE OF WAGE BETWEEN MAXIMUM ABOUT 1873 AND 1885.

(Tabulation as in III. *A.*)

CHAPTER VII.

THE GRAPHIC METHOD.

CHAPTER VII.

THE GRAPHIC METHOD.

1. General Purpose.

The two main methods of elementary statistics which ought to be understood by all students or officials who handle figures, which are easily within the grasp of all independently of mathematical training, but are generally misunderstood or ignored by the uninterested or the uninitiated, are the method of averages and the method of diagrams or the graphic method. These two are placed together because the uses of averages and diagrams are nearly related. When we deal with large and complex masses of figures we are unable to grasp them in their entirety, however clearly they may be tabulated. Any list of figures—the populations of different towns, the death-rates at successive ages, the wages of many workpeople, the imports for a series of years—becomes less comprehensible as its length increases. A series of ten numbers can, perhaps, be easily grasped, of twenty only with an effort; while a printed list of figures for one hundred successive years leaves hardly any impression on our mind at all; we cannot see the wood for the trees. The test to which all questions as to the use of averages should be referred is that the averages selected should afford the best summary of the whole group in question that the mind can grasp. When the meaning of the word average was sufficiently extended, we found that we could select three, four, or even ten suitable figures which adequately showed the main features of any group. The main use of diagrams is also to present large groups of figures so that they shall be intelligible in their entirety, and the test for all diagrams is that the diagram as drawn should afford the best view of the series or group of figures that the eye can appreciate. Diagrams have one use which averages have not, for it is only by a diagram that a series of figures relating to successive years can be adequately presented; but in reality they are less essential than averages, for

Averages and diagrams.

the latter often have an existence independently of the figures from which they are derived, representing true types of the quantities which are being measured; and by their use alone are further comparisons of complex groups made possible: while diagrams, on the other hand, might be dispensed with, being auxiliary rather than essential, merely an aid to the eye and a means of saving time.

Graphic representation of averages.

To connect this chapter more closely with the preceding, we will show how the same group of figures, for example the wages of a large group of workpeople, may be represented by either method.

Consider the following data:—

Numbers of workpeople earning—

From 15/ to 16/	-	200		From 25/ to 26/	-	1,200	
„ 16/ „ 17/	-	400		„ 26/ „ 27/	-	800	
„ 17/ „ 18/	-	100	1,000	„ 27/ „ 28/	-	700	3,500
„ 18/ „ 19/	-	100		„ 28/ „ 29/	-	500	
„ 19/ „ 20/	-	200		„ 29/ „ 30/	-	300	
„ 20/ „ 21/	-	200		„ 30/ „ 31/	-	300	
„ 21/ „ 22/	-	300		„ 31/ „ 32/	-	400	
„ 22/ „ 23/	-	300	2,200	„ 32/ „ 33/	-	400	2,100
„ 23/ „ 24/	-	500		„ 33/ „ 34/	-	500	
„ 24/ „ 25/	-	900		„ 34/ „ 35/	-	500	

From 35/ to 36/	-	600	
„ 36/ „ 37/	-	400	
„ 37/ „ 38/	-	100	1,200
„ 38/ „ 39/	-	80	
„ 39/ „ 40/	-	20	

Using the method of averages we should replace this group by the following figures:—

	s.	d.
Average of all - - - - - -	27	6
„ lowest 1,000 - - - - - -	17	0
„ highest 1,000 - - - - -	36	6
„ middle 4,000 - - - -	27	0

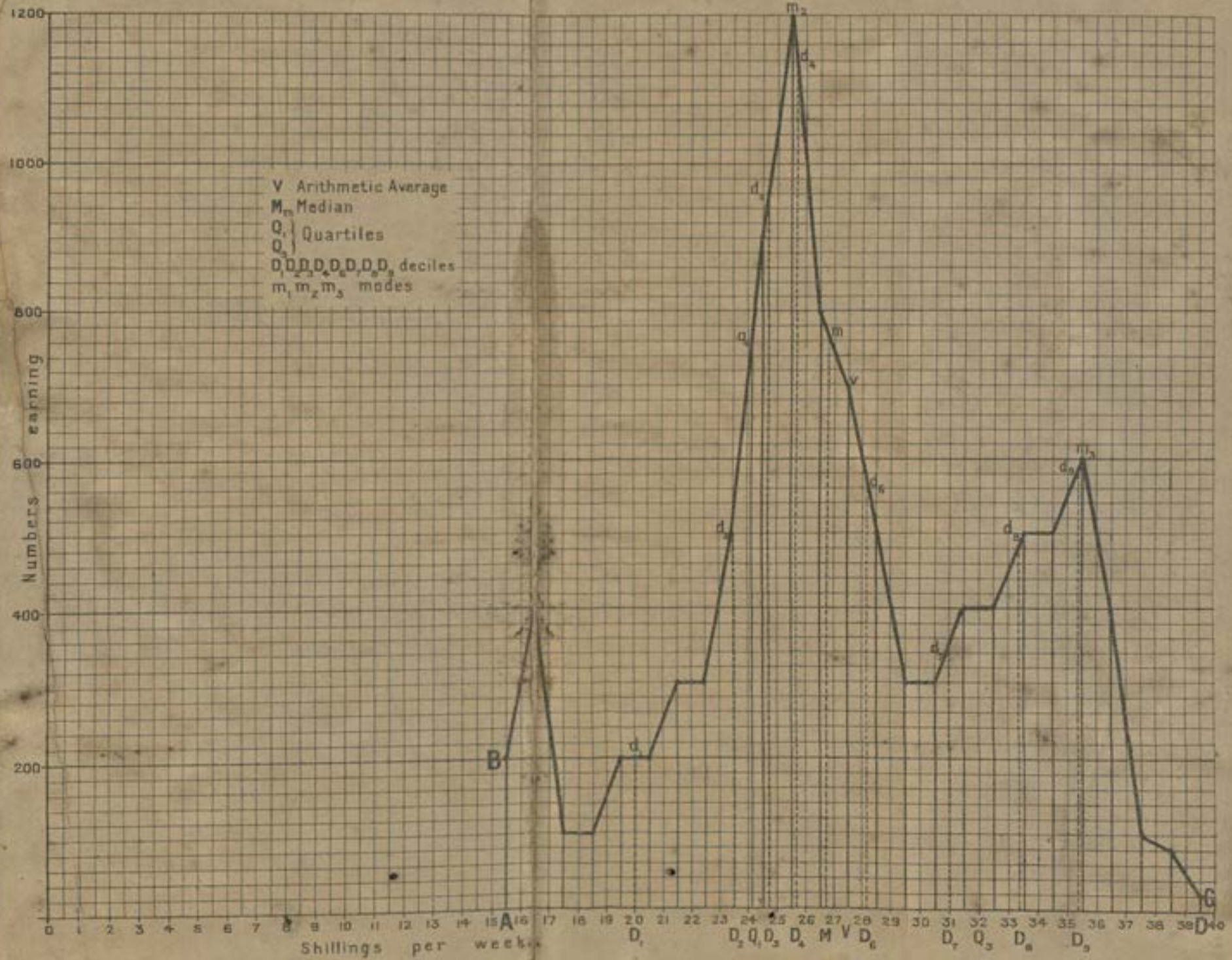
or

Median, 26/9; quartiles, 24/2, 32/.
Deciles, 20/, 23/6, 24/9, 25/8, 26/9, 28/2, 31/, 33/4, 35/4.
Mode, 25/3; secondary positions, 16/6, 36/.

or

Persons earning from -	15/ to 20/	20/ to 25/	25/ to 30/	30/ to 35/	35/ to 40/
Percentages of all - -	10	22	35	21	12

GRAPHIC REPRESENTATION OF WAGE STATISTICS.



To face page 145.

This group is represented on the annexed diagram, an example of the graphic representation of the relation between two variable quantities. A figure similar to this may be used to show birth, marriage, or death rates at different years, numbers of persons of various statures, demand at different prices, or any such group of homogeneous quantities. The same construction can be used to show the changing values of any number in a series of years. Draw a line parallel to the bottom of the page, and mark equal intervals to represent a quantity which can have many successive small increments, such as age, income, height, price, time, and so on. This is called the axis of *abscissæ*, and the distance of a point measured from the zero position along the line is called its *abscissa*. At right angles to this line, parallel to the side of the paper, through the zero position we draw another, called the axis of *ordinates*, and grade this to correspond to the numbers possessing the qualities represented by the abscissæ; at each grade on the axis of abscissæ, draw lines at right angles to it, to represent on the chosen scale the numbers at that grade; these lines are called the *ordinates*. In the annexed diagram the abscissæ represent the amounts of wages, the ordinates the number of persons earning them. Join the tops of the ordinates by straight lines and the diagram is complete. In practice, when squared paper is used, without drawing the ordinates their tops can be marked.

Construction of simple diagrams.

This diagram shows at one glance the distribution of the wage-earners according to their wages. A small number earned between 15s. and 16s., a slightly larger group between 16s. and 17s., very few between 17s. and 19s. Above 19s. the number continually rises; high numbers are found from 24s. to 27s., the highest between 25s. and 26s. The line falls to the 30s. group, but not so low as between 17s. and 19s., then it rises regularly to 36s., and falls rapidly to 39s. Here, then, we have the main group congregated in the neighbourhood of 25s., a distinct but smaller group at 36s., and a small and nearly isolated group at 16s.; representing a considerable group of highly-skilled men between 30s. and 40s., the great mass with ordinary skill between 20s. and 30s., and a small group of incompetents at 16s. These features would not be so easily seen from the tabulated figures.

Description of the wage diagram.

It is to be noticed that the number tabulated as between

15s. and 16s. is represented by the ordinate at 15s. 6d., the middle of the interval; if the original figures on which the table was based had been given to the nearest 1d., the ordinate should be drawn at 15s. 5½d.* It is important that these middle points should be accurately placed.

Continuity.

The use of the line joining the tops of the ordinates is twofold. First, it enables the eye to judge relative heights more easily; and secondly, it suggests the idea of continuity, which can be better illustrated by the next diagram. In this the abscissæ represent ages, the ordinates the estimated numbers of persons living at and above the ages at which they stand per thousand inhabitants of England and Wales at the middle of the year 1891. The ordinates were drawn at the points on the axis of abscissæ representing the middle of each year of age; but length of life cannot be expressed exactly in years, or even in months, days, or minutes. The intention of the diagram is to show the proportion living above each age, and for this purpose the joining line should have no breaks or sharp angles, but should suggest absolute continuity.

In practice, it is useless to mark in the points for smaller intervals than a year, for the eye could not grasp the detail. It is, however, implied that the line drawn has the same shape as that which would result if the number of persons was infinite and the subdivision by age infinitesimal.

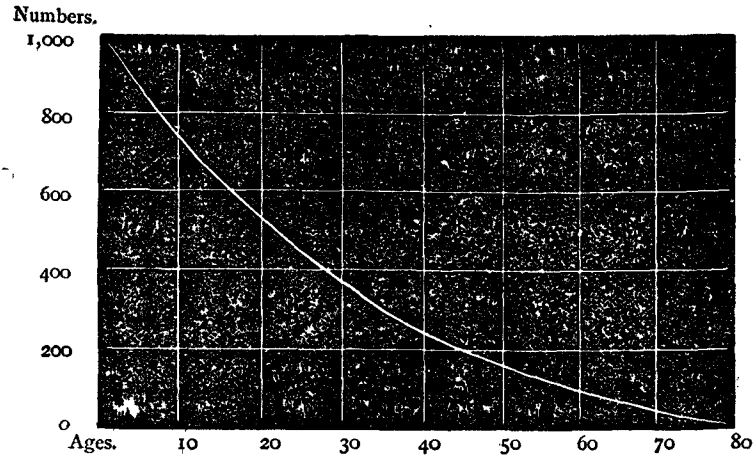
Estimated number per 1,000 of the population at and above—

Ages.		Ages.		Ages.		Ages.		Ages.	
0	1,000	16	628	32	346	49	152	65	47
1	973	17	607	33	332	50	143	66	43
2	949	18	587	34	318	51	135	67	38
3	925	19	567	35	305	52	127	68	34
4	901	20	547	36	292	53	119	69	31
5	877	21	528	37	280	54	112	70	27
6	854	22	510	38	268	55	104	71	24
7	830	23	491	39	256	56	98	72	21
8	807	24	474	40	244	57	91	73	18
9	783	25	456	41	233	58	85	74	15
10	760	26	439	42	222	59	79	75	13
11	738	27	423	43	211	60	73	76	11
12	715	28	407	44	201	61	67	77	9
13	693	29	391	45	191	62	62	78	8
14	671	30	376	46	181	63	57	79	6
15	649	31	361	47	171	64	52	80	5
				48	161				

Calculated from the Census of 1891.

* See p. 88, *supra*.

NUMBERS PER 1,000 OF THE POPULATION ABOVE ASSIGNED AGES.



Apply these remarks to the diagram facing p. 145. Average earnings for a year will not be reckoned exactly by shillings or even pence; if we had a sufficient number of instances we should get regular sequences of earners at successive farthings, and the line representing them would have no sharp angles, but be continually curved. The figure rightly gives the eye this impression of continuousness. Similarly in the diagram representing exports facing p. 151, the line correctly gives the impression that exports are continuous day by day.

By an obvious step we may suppose that the unit of *area*, that contained between vertical lines through two consecutive divisions on the axis of abscissa, and horizontal lines through two consecutive divisions on the axis of ordinates, represents one wage-earner, and it is then easy to see that the area contained between the base line, the curve, and two vertical lines through the points marking any two amounts of wage represents the total number earning rates between those amounts.

Area.

Hence the lines (see p. 145) through M, the position of the median, Q_1, Q_3 those of the quartiles, D_1, D_2, D_3, D_4, D_6, D_7, D_8, D_9 of the deciles divide the area AB$m_1m_2m_3$CD into two, four, and ten equal areas respectively. The centre of gravity of this figure lies on the vertical line through V, the average wage; and the feet of the ordinates through the highest points m_1, m_2, m_3 are at the modes.

The details of technique of diagram drawing, the position of the scales, the devices for making the figure clear, and so on, can be gathered from the various diagrams given in this chapter. The degree of accuracy to which the figures should be marked, whether correct to a million, a thousand, or a unit, is determined simply by the power of the eye to grasp detail; in most of those here given it will be found that a displacement of one in a thousand is perceptible, and this is the ordinary limit. More minute accuracy is useless, for it is not the function of diagrams to dispense with lists of numbers, but only to enable the eye to perceive their significant features.

Requisite accuracy.

Before discussing the choice of scales on which the numbers are to be represented, it is necessary to consider the ways in which a diagram makes an impression on the eye. The eye can judge—(1) Distances; (2) ratios; (3) angles. The dotted lines in the diagram facing p. 151 will illustrate these points. (1.) The eye is a fairly safe judge of distances; there is very little doubt which of two points is the further from the base line; when squared paper is used, a difference of 1 in 1,000 is perceptible. The eye can also judge differences quickly. In the figure the value of the exports in 1883 exceeded that in 1885 by more than the value in 1890 exceeded that in 1883. (2.) It can be quickly seen that the value of exports doubled between 1862 and 1889; or that the value in 1878 is three-quarters of that in 1890. The accuracy with which the eye can make such measurements is not great; it is not easy to detect that the ratio of the values in 1873 and 1871 (1.095 : 1) is greater than the ratio of the values in 1882 and 1880 (1.073 : 1); but the general impression given by the diagram is partly made up by unconscious calculations of this nature. To make these observations accurately the method described on pp. 188-9 should be used. Notice that for these observations the insertion of the base line is necessary; and, because they are made unconsciously, there are very few cases where a diagram without a base line gives a correct impression. (3.) The question, Was the increment greater in 1887-88 or in 1888-89? can be more quickly answered by observing the angles than by noting the differences. The line showing the latter change is steeper (makes a greater angle with the horizontal) than the line showing the former. Hence the latter increase is the greater; actually £14,400,000 against

Optical power.

£12,600,000. The most useful exercise of this power, however, is to judge the dates at which the rate of increase changed; thus the value of exports increased in 1862-63, increased at a slower rate in 1863-64, and slower yet in 1864-65, more rapidly in 1865-66; a slow fall followed in 1866-67, then an increase began which is continually accelerated to 1871, and so on. The line from 1872-76 is concave to the base line, showing an accelerated fall; the concavity from 1879 to 1882 corresponds to a retarded rise; at 1888 convexity gives place to concavity, for at that date the rate of increase began to diminish.

It is difficult to lay down rules for the proper choice of the scales by which the figure should be plotted out. It is only the ratio between the horizontal and vertical scales that need be considered. The figure must be sufficiently small for the whole of it to be visible at once; if the figure is complicated, relating to a long series of years and varying numbers, minute accuracy must be sacrificed to this consideration. Supposing the horizontal scale decided, the vertical scale must be chosen so that the part of the line which shows the greatest rate of increase is well inclined to the vertical, which can be managed by making the scale sufficiently small; and, on the other hand, all important fluctuations must be clearly visible, for which the scale may need to be increased. Any scale which satisfies both these conditions will fulfil its purpose. The annexed page shows the erroneous impressions which can be given by a judicious manipulation of the scale and by the omission of the base line. The diagrams, which are drawn roughly, all represent the same estimates of wages in England and in the United States of America for certain years from 1860. Figure 1 sets the lines in proper relief. In figure 2, the base line is not drawn in the zero position for the English scale, and the American scale is reduced; the consequence is that English wages appear to have fluctuated widely, while American made steady progress. In figures 3, 4, and 5 the scales are doctored and the base line adjusted, so that in 3 American wages seem to have caught up English, in 5 exactly the reverse is the case, while in 4 wages appear to have moved with equal rapidity in both countries. An examination of these figures will show that the eye cannot be trusted to supply the right base line, or to estimate the importance of fluctuations without it; and, with

Choice of scale.

Necessity of correct base line.

The Same Figures Represented on Various Scales and with Erroneous Base Lines.

In each figure the scale for English wages is on the left, that for American wages on the right.

——— *English Wages.*
..... *American Wages.*

Figure 1.

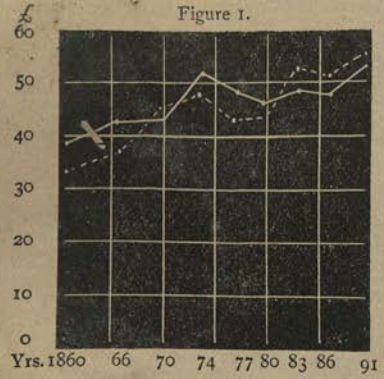


Figure 2.

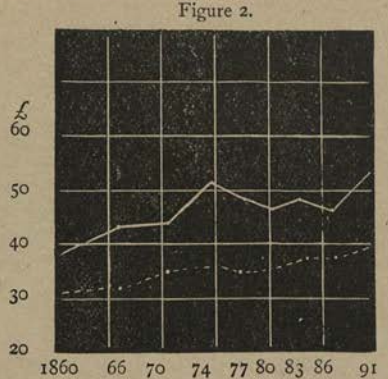


Figure 3.

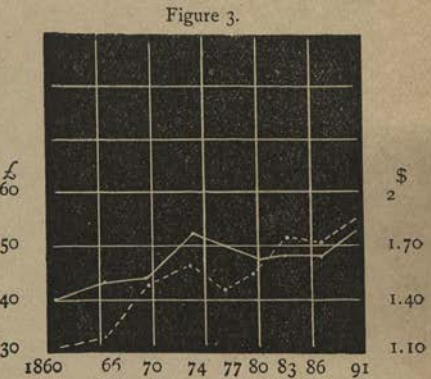


Figure 4.

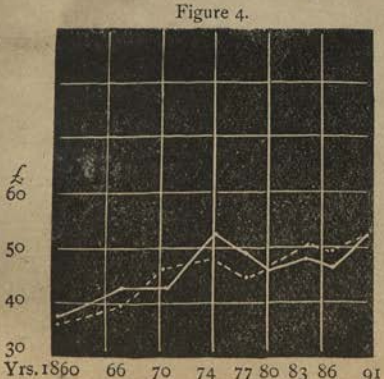
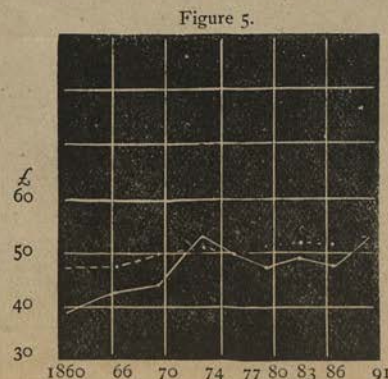


Figure 5.

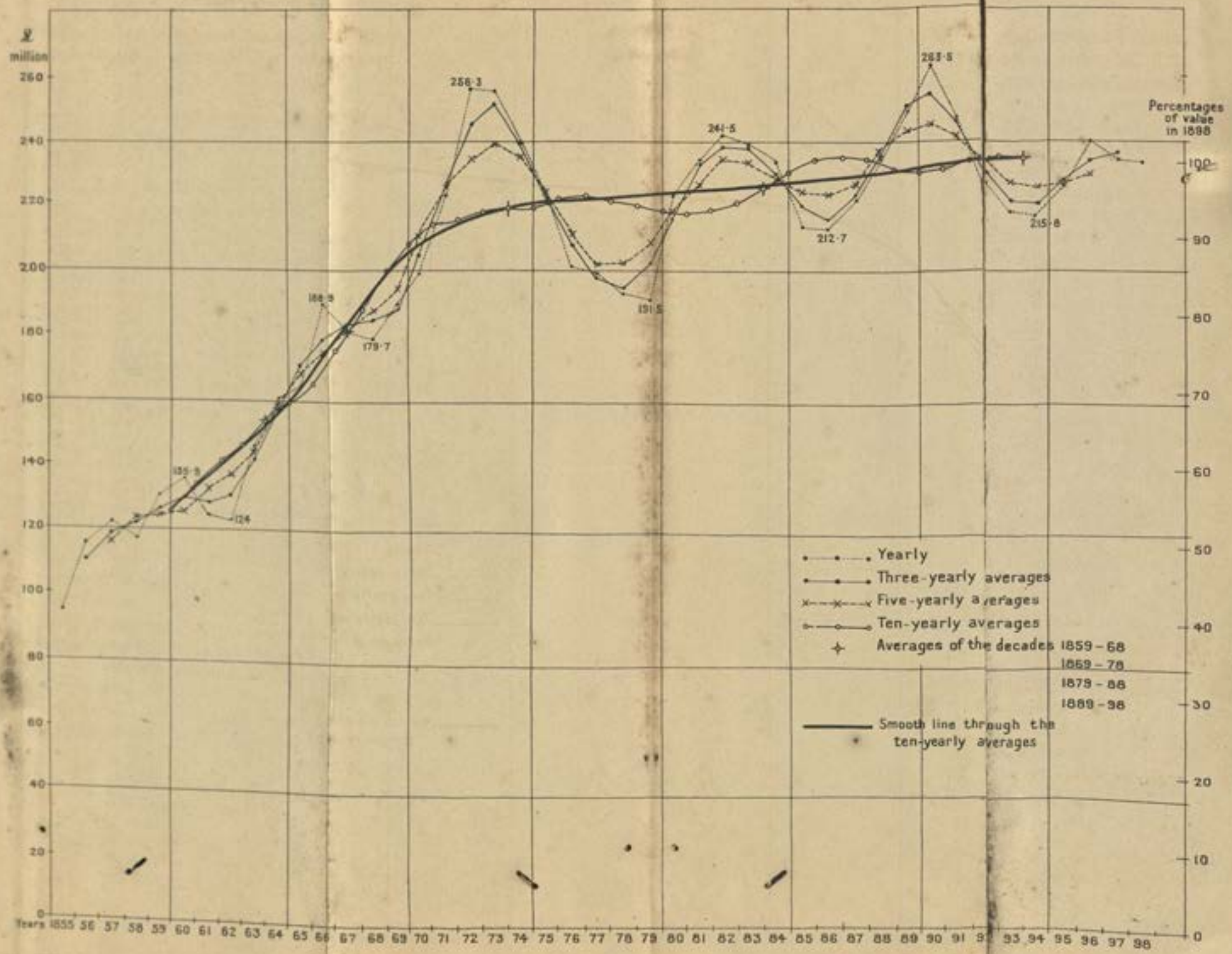


Average Wages

Years	In England per Annum.	In U.S.A. per Diem.
1860	£38.8	$1.10
1866	43.2	1.22
1870	43.9	1.50
1874	52.4	1.60
1877	49.8	1.48
1880	47.6	1.57
1883	48.8	1.75
1886	48.0	1.71
1891	53.2	1.85

TOTAL VALUE OF BRITISH AND IRISH PRODUCE EXPORTED
FROM THE UNITED KINGDOM, 1855-1898.

Example of the Method of smoothing Curves.



To face page 151.

certain exceptions to be mentioned later,* it is well to distrust all those numerous diagrams, where space has been economised at the expense of the base line.

TOTAL DECLARED REAL VALUE OF BRITISH AND IRISH PRODUCE EXPORTED FROM THE UNITED KINGDOM. 1 = £1,000,000.

		Averages.					Averages.		
		Three Yearly.	Five Yearly.	Ten Yearly.			Three Yearly.	Five Yearly.	Ten Yearly.
1855	95.7	...	...	...	1878	192.8	197.4	210.9	218.0
1856	115.8	...	...	...	1879	191.5	194.4	201.4	218.1
1857	122.0	111.2	...	...	1880	223.1	202.5	201.3	220.5
1858	116.6	118.1	...	...	1881	234.0	216.2	208.2	221.6
1859	130.4	123.0	116.1	...	1882	241.5	232.9	216.7	220.1
1860	135.9	127.6	124.1	...	1883	239.8	238.4	226.0	218.6
1861	125.1	130.5	126.0	...	1884	233.0	238.1	234.3	217.9
1862	124.0	128.3	126.4	...	1885	213.1	228.6	232.3	216.9
1863	146.5	131.9	132.4	...	1886	212.7	219.6	228.0	218.1
1864	160.4	143.7	138.4	127.2	1887	221.9	215.6	224.1	220.4
1865	165.8	157.6	144.4	134.3	1888	234.5	223.0	223.0	224.5
1866	188.9	171.7	157.2	141.6	1889	248.9	235.1	226.2	230.2
1867	181.0	178.6	168.7	147.5	1890	263.5	249.0	236.3	234.2
1868	179.7	183.2	175.1	153.8	1891	247.2	253.2	243.2	235.5
1869	190.0	183.6	181.0	159.8	1892	227.1	245.9	244.2	234.1
1870	199.6	189.8	187.8	165.9	1893	218.1	230.8	240.9	231.9
1871	223.1	204.2	194.6	175.7	1894	215.8	220.3	234.3	230.2
1872	256.3	226.3	209.7	188.9	1895	225.9	219.9	226.8	231.4
1873	255.2	244.9	224.8	200.0	1896	240.1	227.3	225.4	234.1
1874	239.6	250.4	234.7	207.9	1897	234.3	233.4	226.8	235.4
1875	223.5	239.4	239.6	213.7	1898	233.4	235.9	229.8	235.3
1876	200.6	221.0	235.1	214.9	1899	255.4*	241 0	237.8	236.1
1877	198.9	207.7	223.7	216.7					

* Not including the newly reckoned value of ships exported.

We can now pass on to the consideration of the smoothing of curves, for which purpose the question of the "alleged stationariness of our exports," discussed by Sir R. Giffen in his paper before the Royal Statistical Society in 1899, affords an excellent illustration. The thin dotted line on the diagram opposite shows the value of exports year by year, and the first impression given by it is that exports have not grown in value in recent years. Sir Robert Giffen gave the following table:—

Smoothing curves.

AVERAGE ANNUAL VALUE OF EXPORTS.

1855-57	£134,000,000
1865-67	228,000,000
1875-77	264,000,000
1885-87	274,000,000
1895-97	292,000,000

See pp. 188-194, *infra*.

and from this he deduced "that all through there is an increase, and that the only sign of stationariness is an increase at a less rate in the last periods than in the earlier periods."

The *Saturday Review* * wrote "that such a conclusion is grossly misleading," for the figures are merely triennial averages of selected years showing a happy coincidence; "why was not 1898 included?" An inspection of the numbers does not show us the answer to this criticism, but on the diagram the whole circumstances are visible at a glance. Since 1865 three great waves have been completed. The maximum of 1872, due to the inflated prices of that year, is very high, but that of 1890 is greater than any previous figure, while the maximum in 1882 is comparatively low. The minima increase throughout; those of 1868, 1879, 1886 show a regular progression, which falls off greatly in 1891. In 1894-96 it looked as if another decennial cycle was in progress, but this has been checked. Since the discussion the returns for 1899 show an increase which brings the figure for that year very near the maximum of 1872.

The *Saturday Review* went on to ask why Sir Robert Giffen did not give "proper quinquennial averages," such as—

AVERAGE ANNUAL VALUE OF EXPORTS.

1870-74 - - - - -	£235,000,000
1880-84 - - - - -	234,000,000
1890-94 - - - - -	234,000,000
1898 - - - - -	233,000,000

and it must be granted that this gives an appearance diametrically opposite to that of the previous table.

It is clear that we need some general method of bringing these figures into a form which shall be quite independent of the choice of any special years. The diagram facing page 151 does this. The thick continuous line, lying almost over the dotted line of annual values, shows triennial averages taken yearly, that is the average of each year with those before and after it; this line smooths off the corners without affecting the general appearance. The line of crosses shows quinquennial averages, each year being averaged with the two previous and two subsequent years. The line of circles shows decennial averages; each circle is placed at the centre of the period whose average it represents;

* January 1899, pp. 66, 67.

thus the circle showing the average of the ten years 1875-84 is placed vertically over the line separating the years 1879 and 1880.*

On looking at the line of quinquennial averages it is clear that the *Saturday Review* did precisely what it accused Sir Robert Giffen of doing, for years are taken which favour the argument. The quinquennial periods selected for comparison with 1898 are all on the upper parts of the waves, the marks showing these averages are very near the maxima of the quinquennial line, while the year 1898 does not appear to be a maximum. We might with just as much or as little accuracy give the following :—

Choice of periods.

QUINQUENNIAL AVERAGES OF THE VALUES OF EXPORTS.

1865-69	£181,000,000
1875-79	201,000,000
1885-89	226,000,000
1898	233,000,000

and say that the value in 1898 was higher than any of the previous selected averages. There is no need to use arbitrary dates to get at the facts. No argument can stand which does not take account of the cycle of trade, which is not eliminated till we take decennial averages. Special marks in the diagram show the averages for 1859-68, 1869-78, 1879-88, 1889-98, and indicate a rapid increase before 1870, and a steady slower progress since. The complete line gives just the same general appearance. If, finally, the figures were completely smoothed by a freehand line keeping as close to this as was possible, without making sudden changes of curvature, the same appearance would be given; the thick line on the diagram is an attempt to do this. The smoothing is obtained by the assumption that the cycle of trade is ten years; when two maxima fall within the same ten years the average of this period by our construction gives the appearance of a maximum (*e.g.*, in 1887) at a date of a minimum. This would be avoided if we continually changed our period for averaging to accommodate the changing wave-length, a somewhat arbitrary proceeding. The difficulty thus arising can be easily corrected by the eye, and the final smoothed line is intended to convey this corrected impression.

* In all the curves of averages the mark showing the average is placed at the centre of gravity of the marks showing the 3, 5, or 10 quantities averaged.

It should be clear now that it was in 1899 five years too soon to pay attention to the particular figure for 1898; the figures for the next five years, necessary to determine the character of the coming wave, could not be foretold. When the figure for 1899 (not represented on the diagram) is included, the new decennial average (1890-99) is the highest on record, while the actual value for the year 1899 has only been exceeded in 1899 and 1873. It will be seen, moreover, that the sentence quoted from Sir Robert Giffen on p. 152 is fully justified.

The smoothed line now constructed represents the general tendency of the value of exports, when accidental and temporary variations are removed. If it were possible to separate entirely variations of short period from secular changes, to separate the ebb and flow of the tide of commerce from the steady current of increasing trade, we may suppose that we should obtain a result represented by this line. In it there are no sudden changes even in rates of growth, while the addition and subtraction year by year of relatively small quantities would produce precisely that irregular fluctuating line from which the smooth line was obtained.

Meaning of smooth line.

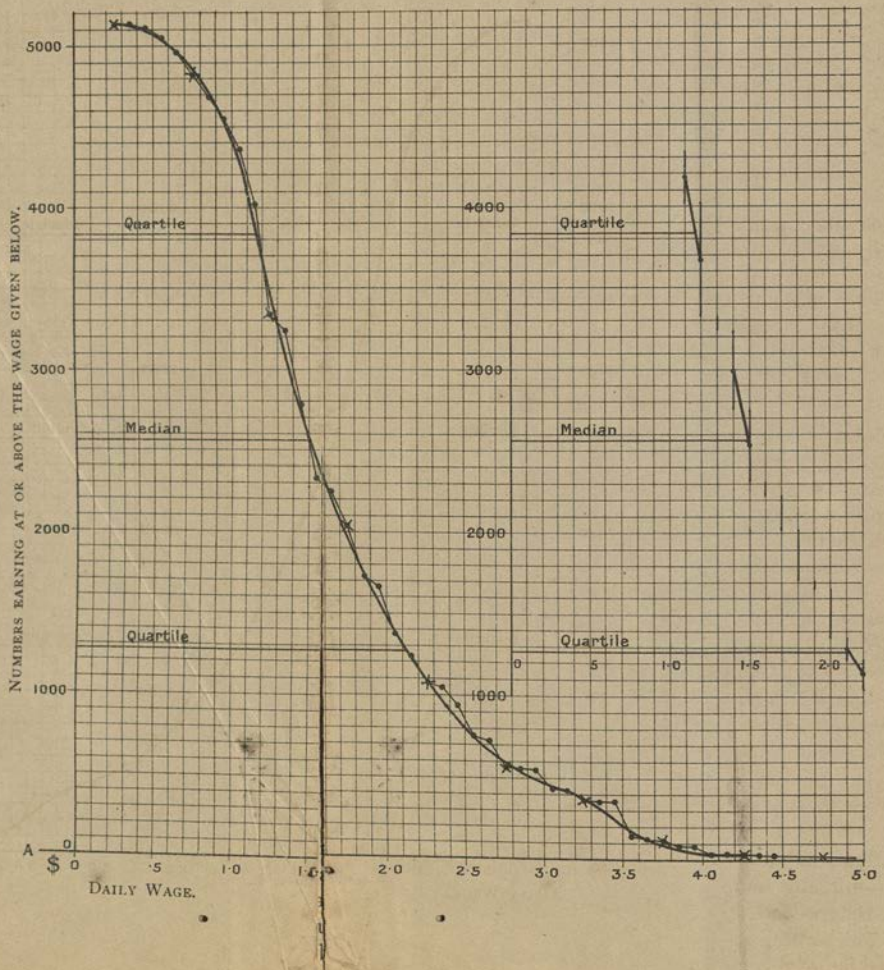
The fuller discussion of "smoothing" series of figures belongs to the chapter on interpolation, but one other group may here be considered, as showing the use of the graphic method for obtaining regularity out of irregular raw material. Referring back to the figures given on p. 120, the wages of 5,000 workers can be expressed anew by a diagram, in which the ordinates represent the numbers earning *at or above* a certain wage. The thin angular line on the adjacent page represents these numbers, entered for every 10-cent group. This plan is especially useful for irregular figures, like this wage-group, for the line must always tend upwards from the numbers earning the highest wage to the numbers earning at least the lowest. The diagram is also at once adaptable to the graphic method of finding the median described on p. 127.

Smoothing a homogeneous group.

The irregularities shown by the thin line do not arise from any law of wage-grouping, but are due to the accidents of observation; if we regard these returns as samples out of a much larger unregistered group, we may suppose that a smoothed curve will indicate approximately the form which would be obtained, if our returns were complete. To smooth this figure,

GRAPHIC METHODS OF DETERMINING THE MEDIAN AND MODES.

For Explanation see p. 154-6.



To face page 155.

draw a freehand line passing as near the points as possible without abrupt changes of curvature, as in the annexed diagram. A new approximation may be made for the median, quartiles, &c., by drawing horizontal lines through the points on the vertical scale corresponding to half, one-quarter, three-quarters, &c., of the workers; from the points where these cross the smooth line, draw vertical lines to the scale of dollars; the points on the scale so obtained are the median (quartile, &c.) wage.

Graphic method of finding the median.

The results obtained are:—

	Median.	Quartile.	Quartile.
Given on p. 92 - -	$1.49	...	...
By method of p. 128, used in annexed diagram -	$1.49	$1.16	$2.12
From smooth curve in annexed diagram - -	$1.51	$1.15	$2.13
By method of interpolation, p. 253 - - -	$1.536	...	...

This method is not, however, one of great precision; a very slight change in the curvature of the smoothed line would make more difference than those shown between the second and third lines in the above table.

This method is more useful for determining the mode. It will be remembered that the difficulties in doing this before arose from the uneven distribution on the two sides of the mode, and in the displacement of the mode by the adoption of a second system of tabulation. The first of these difficulties entirely disappears in the graphic method, while the second is diminished, for the displacement now only depends on the slight possible variations in the curvature of the smooth line. The mode is clearly the position where the greatest number is added, in the present method of representing the figures: that is, the mode is where the line, angular or smooth, is steepest. On the smooth curve the maximum steepness is where the tangent crosses the curve,—in mathematical language, at a point of inflexion. This can be determined mechanically by placing a ruler to touch the curve, and turning it round the curve till it crosses it. On the annexed figure this occurs in the interval between $1.10 to $1.40.

Graphic method of finding the mode.

A more complex method of determining both mode and median, is discussed in Chap. X.

This graphic way of finding these averages has two great advantages. It can be applied to numbers which are given at irregular intervals of graduation (*e.g.*, 30 at 30s. 6d., 40 at 30s. 8½d., 35 at 40s. 1d., &c.) as easily and by exactly the same construction as to more regular returns; and if the smooth curve is carefully drawn, the *number* of modes can be seen at a glance and the individual importance of each can be estimated. In the annexed diagram, the curve is concave to the base line from \$.30 to about \$1.20, convex from about \$1.20 to \$3.15, concave till \$3.40, and then convex till the end. The points of inflexion or the modes are where concavity gives way to convexity. Hence there are two modes, of which that near \$3.4 is of the less importance. The mathematical method of pp. 252-4 shows them to be at \$1.10 and \$3.20; the slight divergence between this result and that of p. 121 illustrates the difficulty of locating the mode exactly.

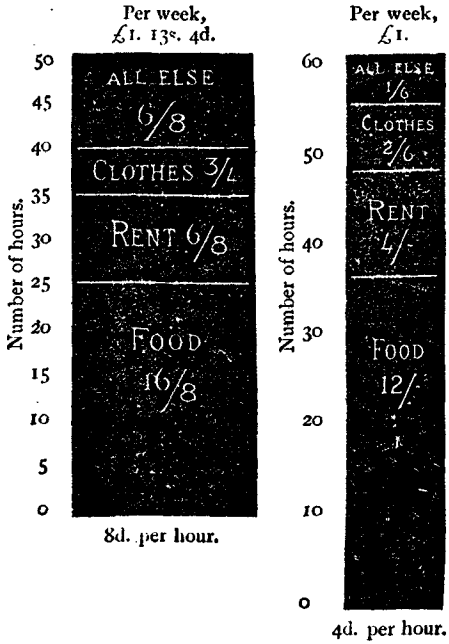
Pictorial diagrams.

A large class of diagrams may be passed by with a few words. Writers and lecturers frequently use points, lines, triangles, squares, circles, even pictures, of different sizes to assist the presentation of the relative magnitude of numbers. These have their use for popular lectures and hand-books, but do not add anything to the significance of the figures. Collections of these may be found in the second volume of Gabaglio's *Teoria Generale della Statistica*, and in M. Levasseur's *La Statistique Graphique* in the *Jubilee Volume* of the Royal Statistical Society.

Of these one group may be signalled as of practical use. Rectangles may be used to express three quantities: one side to represent price; the adjacent side, quantity; and the area, value: or number of houses, average number of inmates and population: or number of hours' work per week, average output or hourly wage, and total output or weekly wage. The figures on the annexed page show the limit to which this method can be usefully pushed.

REPRESENTATION OF THREE FACTS BY RECTANGLES.

Imaginary budgets of an artisan and a labourer, showing amounts spent weekly on various commodities, and number of hours' work necessary for each amount.



The horizontal scale represents pence per hour. .125 inch = 1d.

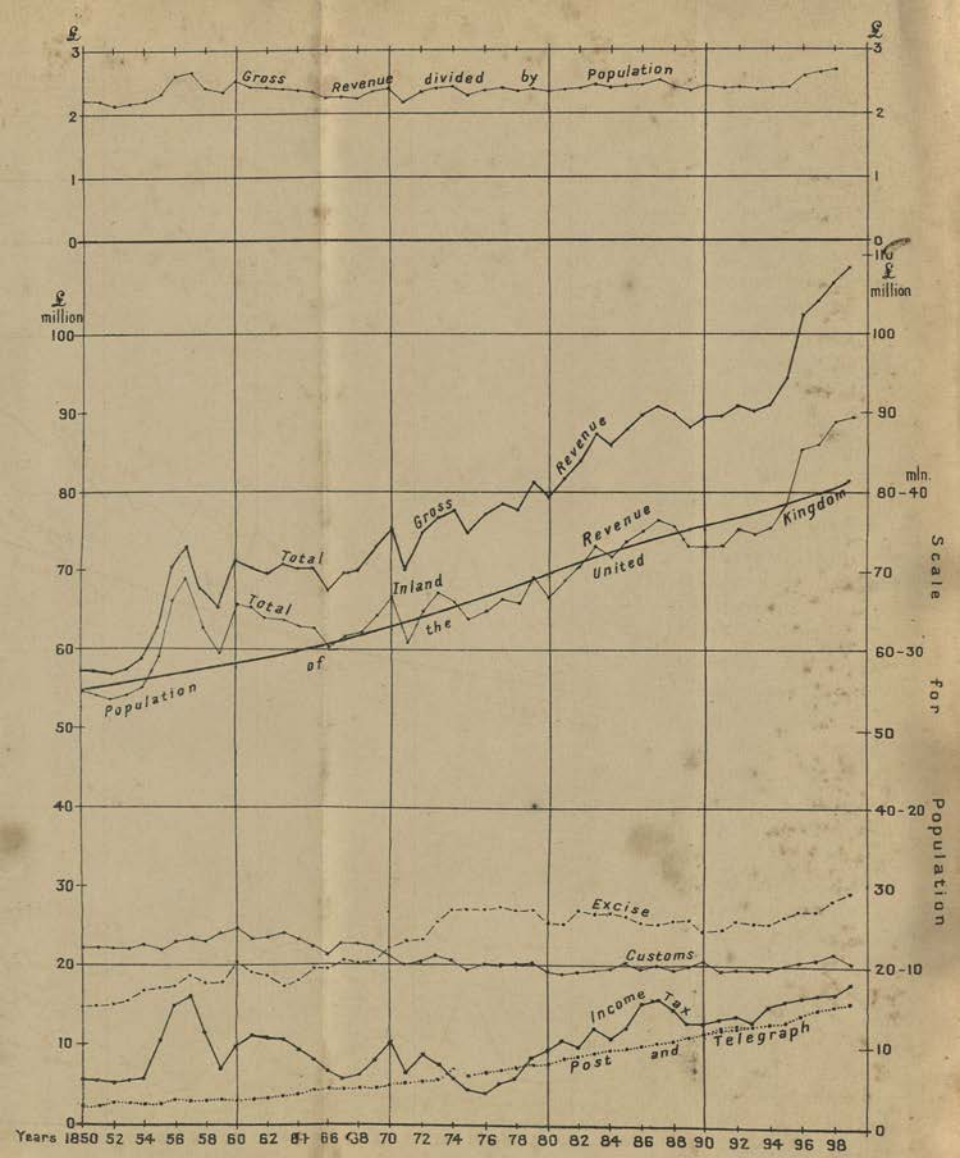
The vertical scale represents number of hours per week. .1 inch = 2 hours.

The areas represent amounts spent, and the whole rectangles show the week's wages on the same scale. 1 sq. in. = 13s. 4d.

Cartograms.

The use of statistical maps needs only a brief notice. Any numerical quality of a population, its density, average income, average taxation, may be shown district by district by suitable markings, or colours. Of these the most useful method is to choose one colour, say blue, for excess above the average; another, say red, for defect. Divide the districts in nine groups, say more than 7 per cent., 5 to 7 per cent., 3 to 5 per cent., 1 to 3 per cent. above the average: these should be marked by four shades of blue, becoming lighter as the average is approached; within 1 per cent. of the average, above or below, should be white; and shades of red, gradually becoming darker, will show the remaining grades below the average. Care must be taken not to adopt too many grades.

For examples of this method see Booth's *Life and Labour of the People*, maps; the *Statistical Atlas of the XIth Census* of the United States; the *Statistical Atlas of India*; and the maps in M. Levasseur's paper just mentioned. A cheap and very effective method, by which similar results are obtained in black and white only, may be seen on Plate I[P] (misprinted 2) in that paper, and in the excellent chapter on Graphic Representation in Bertillon's *Cours élémentaire de Statistique*, p. 133 *seq.*



To face page 159.

2. HISTORICAL DIAGRAMS.

Perhaps the chief use of diagrams is to afford a rapid view of the relations between two series of events.

The different cases that occur are best illustrated by examples. The simplest is when we wish to compare two sets of figures expressed in the same unit, say £ sterling; and the simplest of these when we wish simply to compare a whole and its parts.

Comparison of figures

On the adjacent diagram the upper line shows the annual total gross revenue of the United Kingdom (*Statistical Abstract*, p. 9); the next line, that part which comes from inland revenue and customs, the difference being mainly composed of post office receipts. The principal heads of revenue are customs, excise, income tax, and post office. These are shown by suitable lines for each year, each line being independent of the other, and all having the same base line and being on the same scale. This method is greatly preferable to the alternative one of drawing a second line representing the total less customs, a third the total less customs and excise, and so on, because the eye is then quite incapable of judging the relative movements of the separate items. The figure shows at once the main features of the course of revenue. The increase has been rapid but irregular. The growth in the Crimean War was too rapid to be at once maintained, but the figures for the 60's are at a far higher level than those for the 50's. A rapid fluctuation in 1870 is followed by a more regular growth almost unchecked till 1887; and then, after a short stationary period, there is a great increase in 1895. These remarks apply almost without alteration to the line showing inland revenue and customs. If we look for the parts of the revenue that have borne the increase and change, we see that in the whole period receipts from excise have increased most, next those from the post office, and next those from the income tax, while the customs have diminished. Each line has its distinctive features. The post office payments show an almost regular growth. The income tax fluctuates violently, bearing the brunt of nearly all the rapid changes in the total, especially in 1856

Illustrated by the revenue.

Revenue of the United Kingdom.

Unit, in all columns, £10,000.

	Total Revenue.	Inland Revenue and Customs.	Customs.	Excise.	Property and Income Tax.	Post and Telegraph
1850	5,739	5,431	2,226	1,497	560*	216
1851	5,732	5,412	2,204	1,528	560*	228
1852	5,658	5,335	2,222	1,538	550*	237
1853	5,753	5,401	2,214	1,575	570*	237
1854	5,890	5,502	2,251	1,630	580*	252
1855	6,282	5,944	2,163	1,680*	1,070*	237
1856	7,026	6,601	2,324	1,730*	1,520*	281
1857	7,279	6,848	2,353	1,840*	1,620*	292
1858	6,788	6,309	2,311	1,782	1,159	292
1859	6,548	5,987	2,412	1,790	668	320
1860	7,109	6,570	2,446	2,036	960	331
1861	7,028	6,514	2,331	1,943	1,092	340
1862	6,986	6,412	2,367	1,833	1,036	351
1863	7,060	6,390	2,403	1,715	1,057	365
1864	7,021	6,306	2,323	1,821	908	381
1865	7,031	6,291	2,257	1,956	796	410
1866	6,781	6,036	2,128	1,979	639	425
1867	6,943	6,156	2,230	2,067	570	447
1868	6,960	6,204	2,265	2,016	618	463
1869	7,259	6,422	2,242	2,046	862	466
1870	7,543	6,708	2,153	2,176	1,004	477
1871	6,994	6,106	2,019	2,279	635	527
1872	7,471	6,484	2,033	2,333	908	543
1873	7,661	6,660	2,103	2,578	750	583
1874	7,734	6,608	2,034	2,717	569	700
1875	7,492	6,397	1,929	2,739	431	679
1876	7,713	6,525	2,002	2,763	411	719
1877	7,857	6,636	1,992	2,774	528	730
1878	7,774	6,610	1,997	2,746	582	746
1879	8,115	6,899	2,032	2,740	871	757
1880	7,934	6,695	1,933	2,530	923	777
1881	8,187	6,895	1,918	2,530	1,065	830
1882	8,396	7,058	1,929	2,724	994	863
1883	8,739	7,313	1,966	2,693	1,190	901
1884	8,616	7,187	1,970	2,695	1,072	947
1885	8,799	7,380	2,032	2,660	1,200	966
1886	8,958	7,493	1,983	2,546	1,516	989
1887	9,077	7,611	2,015	2,525	1,590	1,028
1888	8,980	7,566	1,963	2,562	1,444	1,060
1889	8,847	7,360	2,007	2,560	1,270	1,118
1890	8,930	7,341	2,042	2,416	1,277	1,177
1891	8,949	7,358	1,948	2,479	1,325	1,226
1892	9,099	7,534	1,974	2,561	1,381	1,263
1893	9,040	7,480	1,971	2,536	1,347	1,288
1894	9,113	7,543	1,971	2,520	1,520	1,301
1895	9,468	7,865	2,011	2,605	1,560	1,334
1896	10,197	8,512	2,076	2,680	1,610	1,422
1897	10,395	8,597	2,125	2,746	1,665	1,477
1898	10,661	8,855	2,180	2,830	1,725	1,518
1899	10,834	8,945	2,085	2,920	1,800	1,586

* These figures cannot be given accurately within £100,000.

and 1870. The fall in this line in 1872-76 is counterbalanced by the rise in excise; while the excise line shows stationariness till 1870, a sudden jump to 1874, and a very slow growth since that date. Customs, on the other hand, have to some extent taken an opposite course to that of excise, so that the total from the two has not changed very rapidly. There is a very marked stationariness since 1871. At the top of the page a new base line is taken, and the number of pounds per head of the population is shown year by year; it will be seen that the only important increase was between 1851 and 1857, and that since 1860 the fluctuations have been slight.

So far we have found no more difficulty in the choice of scales than previously when dealing with only one line, for all the lines on the larger diagram indicate millions of pounds, and when the unit is £1, a new base line has been adopted. But we may need to show the change of population on the larger diagram. It is necessary, as we have already seen, to use the same base line for the two quantities to be compared; but we may choose any point for the beginning of the new line, adapting our vertical scale, for the eye can judge the proportionate changes wherever the line is placed. It is best to decide this point by defining the problem on which the comparison should throw light. If it is required to compare the growth of revenue with the growth of population since, say, 1850, we should start the new line at the point on the 1850 line where the revenue curve begins, and we can then see how the lines intersect one another again and again. Since 1850, however, is an arbitrary date, this plan lacks definition, and it is more logical to make the lines coincide at the most recent date given, with which any previous date can then be compared. The plan adopted on the diagram given is another alternative; the line is drawn on such a scale that it lies fairly close to that for inland revenue throughout the greater part of its course.

Choice of second scale.

The next diagram, facing p. 164, introduces further difficulties as to the choice of scales. The object of the figure is to show the relations between quantity, value, and price of imported wheat, and population. The line A is first drawn on a scale chosen so as to throw its fluctuations into relief. Population is at once brought into relation with this by calculating the amount per head year by year.

Comparison of quantity and value.

The line C to represent these figures is drawn on a different scale, chosen so that the line shall not cause confusion by continually crossing any of the others on the figure. If the figure was too full this could be treated as on p. 161, the revenue per head. The same scale of years must be used, and for simplicity of calculation and appearance, 100 lbs. consumed per head is measured by the same vertical distance as 10,000,000 cwt. imported. A and C refer to the same quantities, and therefore similar lines are used in both cases. The line B represents value and is shown by a broken line. For this line the choice of scale is more difficult. In the diagrams which follow, instances will be shown where special methods are used to bring out specific comparisons. Here this is not necessary, and a scale is adopted which brings the lines A and B into near relation, and shows the fluctuations of B, while the figure is made simple and intelligible by the representation of £20 by the same vertical distance as 20 cwt.

Details of construction.

The line D shows the changing price of wheat as deduced from columns A and B. The scale is chosen so that it boldly crosses the lines A and B; thus its fluctuations are clearly shown, and the numbers are easily seen, for 2s. per cwt. is represented by the same vertical line as 10,000,000 cwt. If the figure was accurately drawn, lines A and D would lie one over the other in 1876-77; they are therefore shifted very slightly horizontally, and clearness is preserved without the general impression being vitiated.

This line B illustrates some very interesting facts. Its chief characteristic is excessive fluctuation; while a smoothed line would show an upward tendency till 1878 and a fall since that date. The fluctuations are the result of a great number of causes: an increasing population, the fact that wheat imported is only complementary to the home product, which is dominated by the English weather, the variation of harvests all over the world, political events, the fall in the value of silver, the development of means of communication and transport, and all the other causes which determine price. Notice how the effects of all these can be traced. A deficient home harvest synchronized with the rise in 1871-73, a world-wide deficiency with the fall in 1880, a good home product with the fall in 1875. The separation of A from C shows the increase of population. The American Civil War is marked in 1865, the general improvement in transport by the rise before 1875, the fall of

Historical facts illustrated.

prices by the fall since 1878. These various causes, however, often tend to neutralise one another.

IMPORTATIONS OF WHEAT AND WHEAT FLOUR, 1862 to 1898.

Year.	A. Total Quantities Imported. Unit, 100,000 cwt.	B. Total Value Imported. Unit, £100,000.	C. Quantity retained per Head of the Population.	D. Average Price of Wheat and Wheat Flour in Shillings per cwt.
			lbs.	
1862	500	286	185	11.44
1863	309	155	112	10.03
1864	288	135	104	9.37
1865	258	124	93	9.61
1866	294	168	104	11.43
1867	391	285	140	14.58
1868	365	249	130	13.64
1869	444	233	156	10.50
1870	369	196	123	10.62
1871	444	268	151	12.07
1872	476	303	163	12.73
1873	516	344	171	13.33
1874	493	309	162	12.53
1875	595	324	197	10.89
1876	519	279	168	10.75
1877	635	407	202	12.82
1878	597	342	187	11.46
1879	730	400	228	10.95
1880	685	393	209	11.47
1881	713	407	217	11.42
1882	808	449	242	11.11
1883	851	438	252	10.30
1884	669	301	192	9.00
1885	823	337	238	8.19
1886	670	261	188	7.79
1887	802	314	224	7.82
1888	804	315	223	7.82
1889	789	311	219	7.88
1890	824	327	226	7.94
1891	895	396	244	8.85
1892	956	371	245	7.76
1893	938	308	248	6.57
1894	967	268	256	5.54
1895	1,073	302	285	5.63
1896	996	309	257	6.21
1897	887	330	228	7.44
1898	944	377	238	7.99

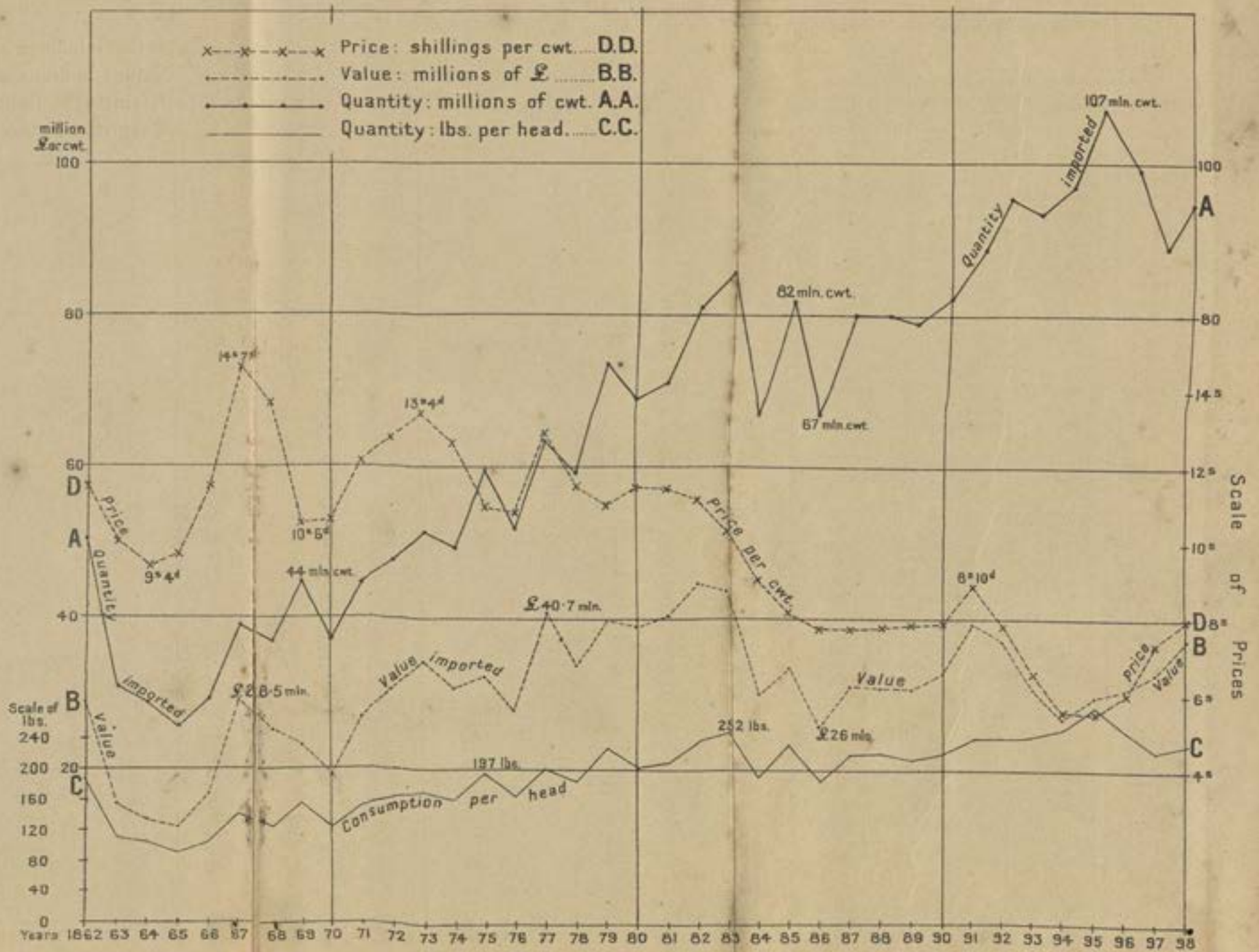
Choice of markings.

As regards the choice of markings for different lines, the chief rule is that lines which cross one another, unless very acutely, must be marked differently. The second rule is to mark similar quantities in similar ways. Thus in the next diagram the lines representing quantities have a resemblance to one another, as have also those showing values; while the two lines relating to imports are distinguished

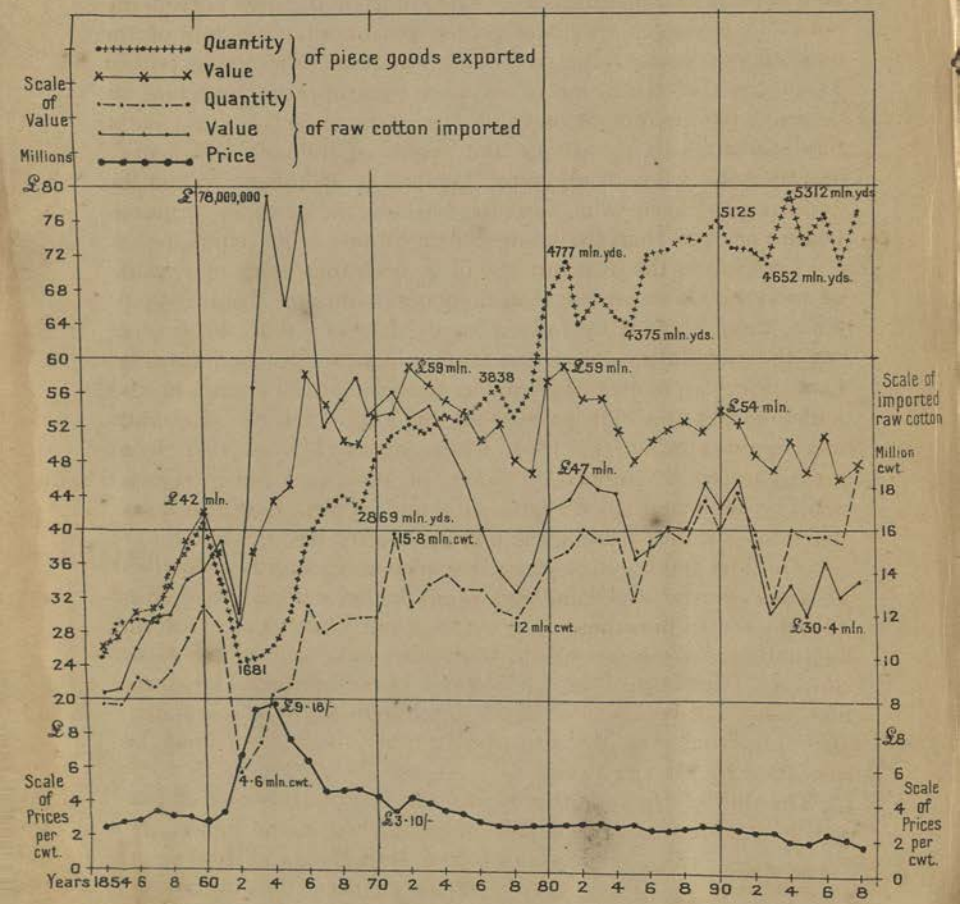
The Cotton Trade, 1854-98.

Year.	Piece Goods Exported.		Raw Cotton Imported.		Price per Cwt.
	Quantity. 000,000's omitted.	Value. 000's omitted.	Quantity. 000's omitted.	Value. 000's omitted.	
	Yards.	£	cwts.	£	£
1854	1,693	25,055	7,923	20,175	2.55
1855	1,938	27,579	7,962	20,849	2.62
1856	2,035	30,204	9,142	26,448	2.89
1857	1,979	30,323	8,655	29,289	3.38
1858	2,324	33,422	9,235	30,107	3.26
1859	2,563	38,744	10,946	34,560	3.16
1860	2,776	42,142	12,419	35,757	2.88
1861	2,563	37,580	11,223	38,653	3.44
1862	1,681	28,562	4,678	31,093	6.65
1863	1,711	37,634	5,983	56,282	9.41
1864	1,752	43,917	7,983	78,219	9.90
1865	2,014	44,876	8,737	66,041	7.56
1866	2,576	57,903	12,299	77,530	6.30
1867	2,832	53,128	11,276	52,003	4.61
1868	2,977	50,265	11,864	55,194	4.65
1869	2,869	49,922	11,907	56,847	4.77
1870	3,267	53,348	11,959	53,478	4.47
1871	3,417	53,643	15,876	55,907	3.52
1872	3,538	58,931	12,579	53,381	4.24
1873	3,484	56,493	13,639	54,705	4.01
1874	3,607	55,023	13,990	50,696	3.62
1875	3,562	53,627	13,325	46,260	3.46
1876	3,669	50,378	13,284	40,181	3.03
1877	3,838	52,442	12,101	35,421	2.93
1878	3,619	48,104	11,968	33,520	2.80
1879	3,725	46,875	13,119	36,181	2.76
1880	4,496	57,678	14,542	42,772	2.94
1881	4,777	59,104	14,992	43,835	2.92
1882	4,349	55,443	15,930	46,655	2.93
1883	4,539	55,534	15,485	45,042	2.91
1884	4,417	51,666	15,618	44,486	2.85
1885	4,375	48,277	12,731	36,473	2.86
1886	4,850	50,172	15,313	38,128	2.49
1887	4,904	51,742	15,995	40,156	2.51
1888	5,038	52,582	15,462	40,009	2.59
1889	5,001	51,388	17,299	45,642	2.64
1890	5,125	54,160	16,013	42,757	2.67
1891	4,912	52,432	17,811	46,081	2.59
1892	4,873	48,766	15,850	37,888	2.39
1893	4,652	47,282	12,650	30,685	2.43
1894	5,312	50,219	15,965	32,944	2.06
1895	5,033	46,759	15,688	30,429	1.94
1896	5,218	51,196	15,669	36,272	2.31
1897	4,792	45,808	15,394	32,195	2.09
1898	5,216	47,910	19,005	34,126	1.80

IMPORTATION OF WHEAT AND WHEAT FLOUR, 1862-1898.



To face page 164.



To face page 165.

from those relating to exports. If it is possible to use more than one colour this principle can be easily carried out.*

The history of the cotton trade.

This diagram is intended to show the relations between the quantities and values of cotton imported and exported during forty years. The vertical scale for values is chosen so as to bring the whole figure to a convenient size and to mark the fluctuations. The value of the raw cotton imported is increased, perhaps trebled by manufacture, and of the finished product a large part is used at home, the rest exported. The excess of the value of exports over imports therefore represents the increment of value due to manufacture (that is the total earnings or the wages and profits of the cotton industry), less the total value of all cotton goods sold at home. When the exports are less in value than the imports, the earnings of manufacture are less than the home consumption: when equal, equal.

Looking at the diagram, it will be seen that value of exports of piece goods exceeded that of general imports from 1854 to 1860, though often by only a small margin; that the reverse was the case during the cotton famine in 1861-66, when extravagant prices were paid for raw cotton to partially supply the home market at a high price, while the export fell off. Equality was again attained in 1867, while since 1871 exports have greatly exceeded imports in value, the difference being permanently established since 1879. It would appear that the home market is saturated, while the foreign market has extended.

The line representing the value exported may be described in a few words: a general and rapid increase took place from 1850 to 1866, interrupted only by the Civil War; since 1866 the fluctuations have been violent, but the general average has fallen slowly. The effect of the Civil War is well emphasised by all the lines here, and is clear also in the diagram facing p. 164. With a little experience in the use of diagrams these lines may be smoothed by the eye alone.

Choice of scale for quantities.

The unit of the quantity of imports is 1,000,000 cwt. of raw cotton; one-tenth of this can be distinguished in the figure. In 1854, 7,993,000 cwt. were imported, and their value was £20,175,000. If we represent 2 cwt. by the same vertical length as £5, as done in the figure, the lines begin

* See *Wages in the Nineteenth Century*, by the present author, diagram p. 90.

at practically the same point. Adopting this scale, we are able to see at once the divergence of quantity from value during the period.

In the year 1891, 17,811,000, cwt. valued at £46,081,000, were imported. The sum would have bought 18,096,000 cwt. at the price of 1854, a difference of only 1½ per cent.; so that it happens in this case that the value and quantity lines are nearly together again in 1891. The actual course of prices is shown by the lowest line on the diagram. In 1862 quantity falls more quickly than value as price rises, and as the supply recovered in 1866 value went up before and more violently than quantity, owing to the high price. In 1869 quantity rose while value fell, but otherwise the lines fluctuate together and continually tend towards each other.

The study of quantity and value in exports is more interesting. It is not obvious what commodity is the best representative of cotton exports. In 1895, 5,000,000,000 yards of piece goods valued at £46,700,000, 812,000 pairs of stockings valued at £220,000, 23,800,000 lbs. of sewing thread valued at £3,000,000, and 250,000,000 lbs. of cotton yarn valued at £9,200,000, were exported. A good plan, perhaps, would be to take so many yards of piece goods as equivalent to so many pounds of yarn, the relative prices being the criterion, and to add these together to determine the quantity; in the figure, however, piece goods only are taken.

History of exports.

In this case there is no simple relation between quantity and value at the first date, and there is no simple method of making the two scales correspond. Having marked the value line on the squared paper in use, it was necessary to draw out a new system for the quantities. In 1854, 1,693,000,000 yards were valued at £25,055,000. Then $\frac{1693}{25}$ yards corresponded to £1, *i.e.*, 67.7 yards to a unit; and each number of yards had to be reduced to this scale. This is done in practice quite easily by a mechanical scale, by which numbers can be automatically reduced in any required ratio. The scale is then entered to the right-hand side of the figure. It is of course not easy to read the exact numbers off the figure, but it can be done with the help of a ruler. To avoid this difficulty, the actual amounts can be entered on the diagram at critical places. But after all it is not the object of the diagram to make it possible to read the

numbers; the object is to show the relative rises and falls, and the steepness, and to allow comparisons of the lines. The figures should be taken from a table. No scale is in reality necessary, except for the process of drawing the lines.

The history of quantity of exports of cotton (and of other textiles) is quite different from that of values. Value fluctuates and shows no rise since 1866. Quantity fluctuates, but not greatly, except at the Civil War, and except in the '60's, and in '80-6 shows a general rise. The smoothed curve would rise throughout.

Quantity and value.

One important cause of this difference is, that, as Sir R. Giffen has pointed out, a large sum should be in reality deducted from export values to allow for the import value of the raw cotton, before any conclusions are drawn as to the progress of British manufacture. Now, as we have seen, the price of raw cotton has fallen very fast during precisely the period (since 1865) that export value has not grown. A greater corresponding deduction should be made in the earlier years than in the later, which would result in a definite, if fluctuating, rise in the period. This would not make values increase so fast as quantities; the difference is due to the general causes of the fall of price of manufactured goods. By looking carefully at the diagram it will be seen that the quantity line approached that of value, when the price was falling in 1866-68, and fell away again with the higher price of 1872; after 1872 the quantity line gets nearer again, and crosses the value line in 1875, when the price was the same as in 1850; since 1875, as prices fell, the divergence has steadily increased.

It must be admitted that a study of the diagram representing these figures leads much more rapidly and safely to many interesting conclusions than the table on p. 164 of the figures themselves.

3. Comparisons of Series of Figures.

A. Before proceeding to the study of the next diagram, it will be well to define more exactly what is our object in comparative studies of figures, and to consider the means at our disposal.

When dealing with two series of similar quantities such as the course of trade or population in two countries, we wish to see the general rate of progress (to be done by smoothing the curve), the years of special increase, the dates of maximum and minimum, in fact to compare the three things that the eye can see—the increase, the rate of increase, and the dates of change of rate of increase. The most obvious way to do this is, to take the same scale and base line for both countries and the same unit of measurement; but this method does not take us all the way. We can judge differences, it is true, and the additions in all the years in both countries, and we can see the highest and lowest points and *dates* of change of rate of increase; but we cannot compare rates of increase. It is not easy to judge ratio, though a rough guess at it is possible. Thus if the trade is very different in magnitude in the two countries, equal absolute increments will mean very different relative increments, and it is difficult to be always on one's guard.

Quæsita in comparisons.

The remedy for this is to alter the arrangement of scales. Make a second figure, in which the unit shall be not a sum of money, but a percentage: let 1 per cent. of England's trade, say in 1850, be the unit for the English line; and 1 per cent. of the trade of Germany, at the same date, for the German line. In other words, express the trade of both countries as percentages of their value in a given year, and draw lines to represent these percentages. Alongside the diagram two or more scales can be placed showing the absolute amounts of the trade of each country. Then the rates of increase will be comparable, equal increments representing equal percentages of the trade of each country; and, in addition, the dates at which either country gained ground relatively to the other can be easily picked out. The question as to whether

Percentage scales.

absolute rates or relative rates should be studied is a very common one in statistics. Sometimes the absolute magnitude should be known, as for instance when we want to estimate the effect of measures which will affect the well-being of special classes, or the trade of special countries; sometimes the relative rate, as when we want to watch the progressive increase of different industries, or to be on our guard as to future competitors. The two studies generally require two different diagrams though they may represent the same numbers.

Absolute or relative progress.

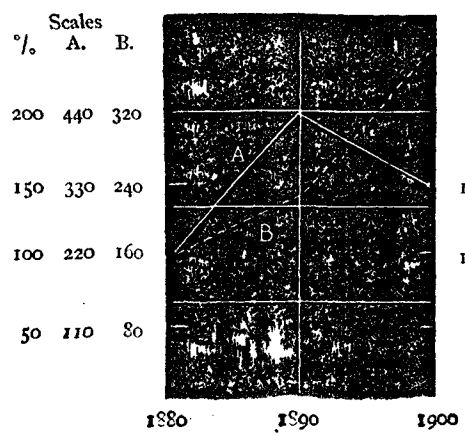
It will be seen that the chief difficulty lies in the choice of the year in which the quantities are to be equated; this must be decided by the nature of the argument which the diagram is to illustrate.

We may compare the following figures—

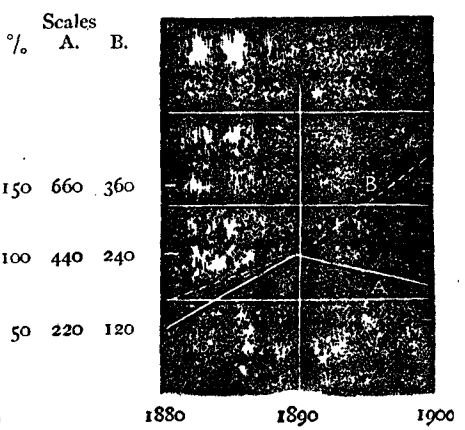
Year -	1880	1890	1900
A - -	220	440	330
B - -	160	240	400

in three ways, thus:—

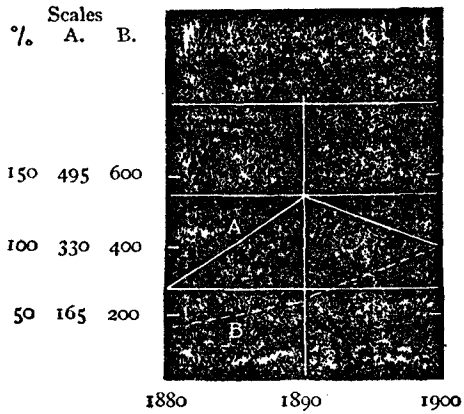
1. Expressed as percentages of values in 1880.



2. Expressed as percentages of values in 1890.



3. Expressed as percentages of values in 1900.



In figure 3 the fluctuations are seen as percentages of the values at the last date, and are thrown into better proportion than in figure 1. It is frequently the case that the equating of quantities at the most recent date throws what are often small beginnings into their right proportion when viewed from the modern standpoint. The statements that the values in 1880 were 40 and 67 per cent. respectively of the corresponding present values, is in better perspective than the statement that the values in 1900 were 250 per cent. and 150 per cent. of the corresponding values in 1880; but circumstances must decide in each case which method is to be adopted.

These points are fully illustrated by the annexed diagrams, the object of which is to analyse the progress of our trade with our colonies and with foreign countries, especially Germany. The first figure shows the total imports and exports, and the parts of each which are colonial and foreign, the scale in millions of pounds being the same for all the lines. A line is also given for imports from Germany, Holland, and Belgium; these are grouped together, because it is not possible to distinguish in the returns from the two latter home manufactures from German goods in transit. It is not clear from this diagram which part of our imports has increased most rapidly. The three lines are, therefore, redrawn in the second diagram, on a percentage scale, all the values being expressed as percentages of the corresponding values in

Illustration from trade with Germany.

1898. It is now seen that imports from foreign countries and from our colonial possessions and India have marched together except during the period of the cotton famine, but the

IMPORTS AND EXPORTS, 1862-1898.

Unit in all columns, £100,000.

	Total Imports.	Total Exports including Re-exports.	Exports to British Possessions.	Exports to Foreign Countries.	Imports from British Possessions.	Imports from Foreign Countries.	Imports from Germany, Holland and Belgium.
1862	2,257	1,662	454	1,207	653	1,604	279
1863	2,489	1,969	550	1,419	847	1,642	283
1864	2,749	2,126	557	1,569	937	1,812	332
1865	2,711	2,188	515	1,673	728	1,982	364
1866	2,953	2,389	572	1,817	722	2,231	388
1867	2,752	2,258	534	1,724	607	2,144	373
1868	2,947	2,278	537	1,741	670	2,277	379
1869	2,955	2,370	519	1,851	704	2,250	405
1870	3,033	2,441	554	1,887	648	2,384	409
1871	3,310	2,836	556	2,280	729	2,581	469
1872	3,547	3,146	656	2,490	794	2,753	455
1873	3,713	3,110	711	2,399	810	2,903	463
1874	3,701	2,977	779	2,197	822	2,879	494
1875	3,739	2,816	767	2,050	844	2,895	515
1876	3,752	2,568	701	1,866	843	2,908	516
1877	3,944	2,523	758	1,766	896	3,049	590
1878	3,688	2,455	720	1,735	779	2,908	575
1879	3,630	2,488	665	1,823	789	2,840	543
1880	4,112	2,864	815	2,049	925	3,187	616
1881	3,970	2,971	867	2,104	915	3,055	582
1882	4,130	3,067	923	2,143	994	3,136	658
1883	4,269	3,054	904	2,150	987	3,282	692
1884	3,900	2,960	883	2,077	958	2,942	646
1885	3,710	2,715	885	1,860	844	2,866	638
1886	3,499	2,690	822	1,867	819	2,680	609
1887	3,622	2,813	823	1,990	838	2,784	646
1888	3,876	2,986	917	2,068	869	3,007	684
1889	4,276	3,156	908	2,248	973	3,304	715
1890	4,207	3,283	945	2,337	962	3,245	694
1891	4,354	3,091	933	2,158	995	3,360	716
1892	4,238	2,916	812	2,104	978	3,260	715
1893	4,047	2,771	786	1,986	918	3,129	720
1894	4,083	2,738	786	1,952	939	3,144	716
1895	4,167	2,858	761	2,098	955	3,211	729
1896	4,418	2,964	907	2,057	932	3,486	761
1897	4,510	2,941	870	2,072	940	3,570	760
1898	4,704	2,940	901	2,039	994	3,709	785

trade from Germany has increased more rapidly than either. If we had equated the quantities in 1862, the German line would have far outpassed the others by 1898; but the impression given would be erroneous as regards absolute quantities, for the

increase was only £50,600,000 for the one, while it was £110,500,000 for the other. The remaining diagram shows the relative rates of increase for Germany, Holland and Belgium, and the British possessions respectively, since 1870.

B. Series of figures are often compared graphically with a view to discovering or illustrating causal relations. In such cases we do not only study relative growth as in the last diagram discussed, but look throughout the period for any signs of resemblance in rates of growth, dates of maxima and minima, or synchronism in any changes. The methods by which such comparisons are made are difficult, and need careful analysis. For instance, we may wish to show that an increase of the allowance for outdoor relief is connected with an increase of pauperism. In this case one line will represent money, the other the number of persons, and there is no common unit; we need not calculate percentages, but having chosen any scale for money, we can make equality in any year by a simple adaptation of the scale for number. We shall wish to establish first, that an increase or decrease of money occurred at, or just before, an increase or decrease in number; and secondly, that the greater the increase of one the greater the increase of the other. In order to show direct connection, we shall try to make one line lie as nearly as possible over the other.

Causal relations.

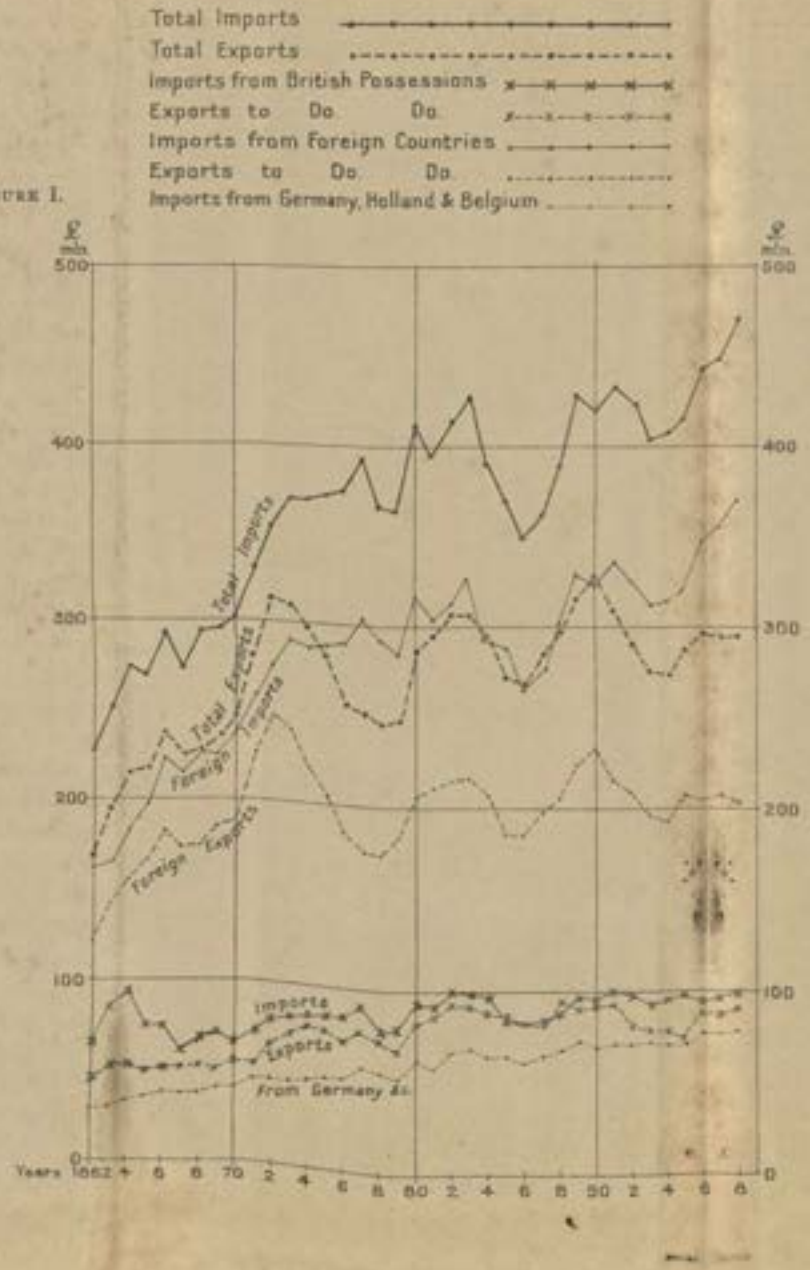
Draw a preliminary diagram in which both lines are entered on any scales; this will suggest the resemblances to be tested. Notice in what period the fluctuations are greatest; this in general should be the period to be taken, for it is here that the causal relations have had most play. If any other period is chosen for any special reasons, these should be made clear, for otherwise a critic may legitimately object that it is only in this period that the connection is distinct. There would be little difficulty in finding short periods in any two curves where the fluctuations synchronised. Take the averages of both money and of number over the period chosen, and draw a second diagram in which the scale for number is chosen by making this average for number equal to the corresponding average for money. Any correspondence between the two lines can be at once detected.

Construction.

The process just described is completely carried out in the first two diagrams comparing the marriage rate and foreign trade facing p. 175.

TRADE with BRITISH POSSESSIONS and FOREIGN COUNTRIES.

FIGURE I.



IMPORTS as PERCENTAGES of THEIR TOTAL VALUES in 1898.

FIGURE II.

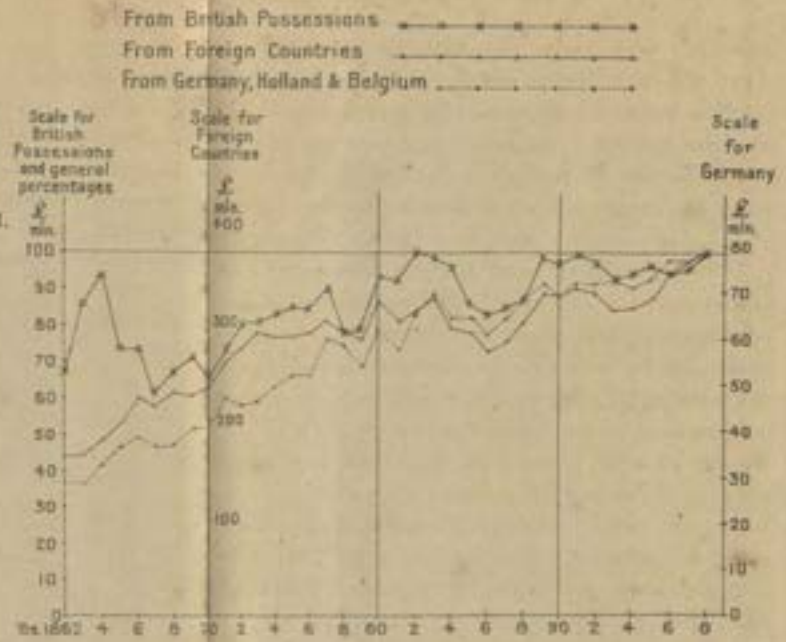

From British Possessions
From Foreign Countries
From Germany, Holland & Belgium
Scale for British Possessions and general percentages
Scale for Foreign Countries
Scale for Germany
£ mln.
Yrs 1862 4 6 8 70 2 4 6 8 80 2 4 6 8 90 2 4 6 8

FIGURE III.

Imports from British Possessions
& Germany, Holland & Belgium
SCALES CHOSEN FOR EQUALITY IN 1870
Scale for British Possessions
Scale for Germany
£ mln.
Years 1870 2 4 6 8 80 2 4 6 8 90 2 4 6 8

To face page 172.

There are many cases when the changes in the magnitudes which we regard as the causes are inversely proportional—in the opposite sense—to those in the magnitudes which we regard as the effects. For instance, if we are comparing trade improvement with the number of unemployed, and make the construction just described, the maxima of the first line would synchronise with the minima of the second. Greater clearness can be obtained by inverting one of the diagrams, plotting out the number employed instead of that unemployed, and then the changes should be in the same sense in both lines.

Inverse relations.

In the above construction the lines will only lie one over the other throughout their fluctuations, if the changes in one quantity are in strict proportion to the changes in the other, if an increase of 10 per cent., for instance, in the allowance for outdoor relief corresponded to one of 10 per cent. in the number of paupers. It is very rare that such a simple relation is found; all we can see in general is that the maxima and minima occur at the same dates, that the fluctuations agree throughout in sense in both series, and that the greater fluctuations in the one correspond to the greater fluctuations in the other.

More complex relations.

Diagrams may often be used to suggest correlation between two series of figures, and this indeed is one of their chief merits, and they may be used to *illustrate* arguments on the subject, but at this point their utility ends, for they cannot be made to *prove* much. Causal relations are very difficult to establish, and the original figures must be critically consulted when theories are to be brought to the test.

Use of diagrams.

We have not yet exhausted the power of diagrams for making such comparisons, but the following method must be applied only with great caution. Suppose that we wish to ascertain whether an increase of 1 bushel in the quantity of wheat to be bought for a sovereign corresponded to an increase of 1.5 in the marriage rate per 1,000, or any such strict numerical proportion. Draw a diagram representing the quantities of wheat, take the average for the period chosen for comparison, and write the scale so as to read 1, 2, 3 . . . bushels *above or below the average. Draw no base line.* Now enter a line to represent the excess or defect of the marriage rate from its average in the chosen period, on a scale such that

More exact method.

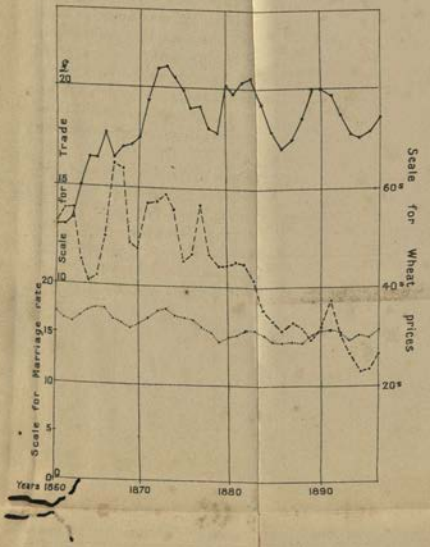
1.5 in excess is represented by the same vertical distance as 1 bushel. The closeness of the two lines would test to what extent the theory was valid. The danger of this method is, that with no base line there is no possibility of judging the amounts of the changes relative to the totals. The insertion of the necessary two base lines would confuse rather than aid.

MARRIAGE RATE, TOTAL EXPORTS AND IMPORTS PER HEAD OF POPULATION, AND AVERAGE PRICE OF WHEAT PER QUARTER.

Year.	Marriage Rate.	Total Exports and Imports per Head.	Average Price of Wheat per Quarter.
		£ s. d.	s. d.
1860	17.1	13 0 8	53 3
1861	16.3	13 0 3	55 4
1862	16.1	13 8 0	55 5
1863	16.8	15 2 7	44 9
1864	17.2	16 8 7	40 2
1865	17.5	16 7 5	41 10
1866	17.5	17 14 5	49 11
1867	16.5	16 9 6	64 5
1868	16.1	17 0 6	63 9
1869	15.9	17 3 9	48 2
1870	16.1	17 10 3	46 10
1871	16.7	19 9 6	56 8
1872	17.4	21 0 0	57 0
1873	17.6	21 4 2	58 8
1874	17.0	20 11 0	55 8
1875	16.7	19 19 4	45 2
1876	16.5	19 0 10	46 2
1877	15.7	19 5 5	56 9
1878	15.2	18 2 1	46 5
1879	14.4	17 16 10	43 10
1880	14.9	20 3 3	44 4
1881	15.1	19 17 5	45 4
1882	15.5	20 8 10	45 1
1883	15.5	20 13 2	41 7
1884	15.1	19 4 1	35 8
1885	14.5	17 16 9	32 10
1886	14.2	17 0 10	31 0
1887	14.4	18 11 7	32 6
1888	14.4	18 12 1	31 10
1889	15.0	19 19 9	29 9
1890	15.5	19 19 7	31 11
1891	15.6	19 14 0	37 0
1892	15.4	18 15 6	30 3
1893	14.7	17 14 9	26 4
1894	15.1	17 11 9	22 10
1895	15.0	17 19 3	23 1
1896	15.8	18 14 1	26 2

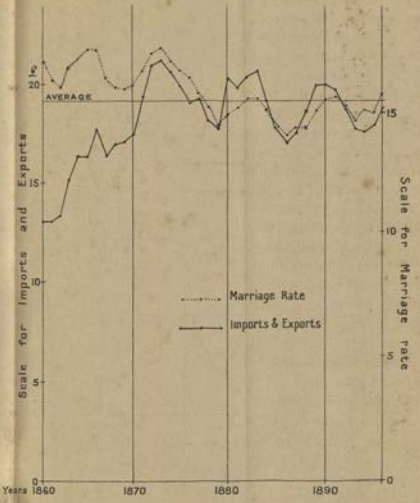
Marriage rate per 1000
Imports & Exports per head
Price of Wheat per quarter

FIGURE I.



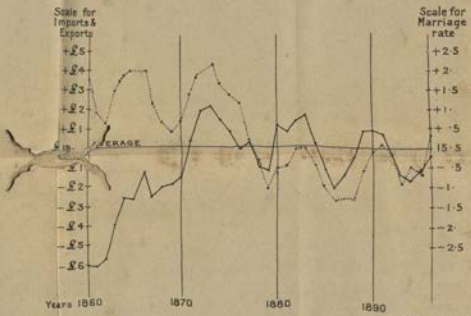
MARRIAGE RATE AND IMPORTS AND EXPORTS PER HEAD.
Averages 1869-94 on same line.
Same base line.

FIGURE II.



MARRIAGE RATE AND IMPORTS AND EXPORTS PER HEAD.
Averages 1869-94 on same line.
Average fluctuations, 1867-95, equal.

FIGURE III.



To face page 175.

It is clear from the preceding analysis that, by the choice of scales and base lines, the points at any two dates may be made to coincide on any number of accurately drawn lines representing series of figures.

The preceding paragraphs are completely illustrated by the adjoining diagram.

On the left are given lines representing the price of wheat in shillings per quarter, the total of values of exports and imports divided by the population, and the marriage rate per 1,000. The scales chosen are simply those which are easiest to use, and throw the lines into proper relief. The points in each scale for the same years are over one another, but the scales differ. The base lines need not coincide.

Illustration of method.

We can see at a glance whether there is resemblance between the courses of these figures. There is at any rate a general correspondence between the fluctuations of trade and of the marriage rate since 1870, and possibly earlier. There are points of likeness between wheat prices and trade; in 1870-73 both rise together, and fall in 1873-75; both rise in 1876-77, fall in the following two years, and then rise again; both fall from 1881 to 1886 and then rise. There are also many cases in which the motions do not agree, especially 1862-64, and 1887-89.

Marriage rate and trade.

If we look now at the price of wheat and the marriage rate, which in the earlier part of the century used to be closely related, the one rising when the other fell, we see that there is no great resemblance either in this or the contrary sense. In 1860-62 and in 1862-64 wheat rose and fell, while the marriage rate fell and rose; wheat rose in 1865-67, while the marriage rate was first stationary and then fell a little; then it continued to fall in 1868-70, though wheat was falling also; in 1870-80 the marriage rate shows one long, wheat two short, fluctuations. Since 1880, in years in which wheat fell, the marriage rate in general fell also and *vice versa*.

Marriage rate and wheat.

Let us consider for a moment the possible links of connection between these phenomena. When wheat was the chief object of expenditure of the working class, its price was the chief thing for them to consider; and so when wheat rose the marriage rate fell. On the other hand, now that wheat is cheap and wages higher, a change in the price of the loaf is only of great importance to a minority;

Connecting links.

it is now the general prosperity of the country, well indicated by the condition of foreign trade, that raises the marriage rate.

When exports and imports are increasing in value, trade is stimulated, and in spite of rising prices, marriageable people are sanguine that the prosperity will remain and the prices fall; but when the prices fall, so do the profits and incomes, and marriageable people are more prudent. For these reasons we may expect the marriage rate and foreign trade lines to resemble each other.

Now the increase of the marriage rate corresponding to an inflation of trade, and an inflation of trade to a time of rising prices in general, we shall find the price of wheat in particular, which is connected with the course of prices in general, rising when trade is inflated and falling when it is depressed, and therefore rising and falling with the marriage rate. But since the price of wheat is influenced also by special causes, it will not always correspond to the state of trade, and still less to the marriage rate, with its former tendency to opposite variations.

There is no need then for surprise that the curves marriage rate and trade correspond; that wheat and trade correspond, but less closely; and that wheat and marriage show a double tendency. The correspondence between marriage and trade is investigated on the diagram. That between wheat and trade should be done on an identical method. Marriage and wheat should be compared twice on different plans: first for direct correspondence, and then by redrawing the wheat curve with its base line at the top for inverse correspondence.

To effect the comparison between the course of trade and the marriage rate, the following steps are taken. On examining the two curves on the first figure, it is seen that the resemblance does not begin before 1869; the parts of the curves since 1869 should therefore be brought into close correspondence. The average marriage rate, 1869-94, is 15.5, and average imports and exports per head, £19. The marriage curve is drawn in the ordinary way; then with the help of a sliding scale the trade curve is put in, so that with the same base line £19 falls on the 15.5 line.

Construction of diagram.

The result is that the curves are seen to rise and fall at the same dates, but not to the same extent; for, while the lines keep nearly parallel from 1873 to 1879, the falls from the maximum being equal, after 1879 the trade line fluctuates further above and below its average than the marriage rate does.

Final Comparison.

It remains to test graphically whether the changes are proportional to one another. An equation of scales may be obtained in the following manner:—

MARRIAGE RATE.

	Maxima.	Minima.	Differences.
1869	...	15.9	
			1.7
1873	17.6	...	
			3.2
1879	...	14.4	
			1.1
1882-3	15.5	...	
			1.3
1886	...	14.2	
			1.4
1891	15.6	...	
			.9
1893	...	14.7	

Average of differences - 1.6

IMPORTS AND EXPORTS PER HEAD.

	Maxima. £ s. d.	Minima. £ s. d.	Differences. £ s. d.
1867	...	16 9 6	
			4 14 8
1873	21 4 2	...	
			3 7 4
1879	...	17 16 10	
			2 16 4
1883	20 13 2	...	
			3 12 4
1886	...	17 0 10	
			2 18 11
1889	19 19 9	...	
			2 8 0
1894	...	17 11 9	

£3 6 3

Represent £3. 6s. 3d. on the same scale as 1.6; then £1 and .5 will be represented by the same height.

This is making the hypothesis that a change of £1 in the total trade per head synchronises with a change of .5 in the marriage rate per thousand. The scales so chosen are marked above and below the common average line in the right-hand figure.

It is now seen that the fluctuations since 1880 lie more closely together in the two curves, but that this closeness has been obtained by the partial sacrifice of the years 1872-80, and there is now a complete disagreement before 1870. A yet shorter period, 1879-1893, would show a very close agreement; but so special a selection would vitiate any general argument.

Our conclusion is, that since 1870 the causes which affect foreign trade have also affected the marriage rate at the same dates and in the same sense, and that the more marked the effects on the one, the more marked are the effects on the other also, but that there is no law of simple proportion between them.

Note.—The relations tested by the middle diagram may be represented by the equation $\frac{x}{a}=\frac{y}{b}$, and that of the right-hand diagram by $\frac{x-a}{y-b}=c$ (a constant), where x and y stand for the value of trade and the marriage rate, and a and b for their average values, and c is chosen so as to make the average fluctuations of the two sets of quantities equal. By the method of least squares c could be chosen so that the correspondence should be closer than with the value given by the calculation in the text.

See note on p. 193 for further methods of making comparisons.

4. Periodic Figures.

We now come to the consideration of periodic figures; that is, of figures which within a given period, in a year for instance when returns are monthly, reach maxima and minima at assigned times, and show fluctuations recurring with regularity in successive periods. In physical phenomena, such as the sunrise, the same daily numbers will represent the phenomena, almost without change, year after year. In the case of the tides we find a link between the more rigid annual curves of seasonal phenomena, and the less marked periods of social statistics; for the tides are subject to separate influences with periods of 24 hours, 24 hours 50 min., 29 days, 1 year, and others, and the effects of these influences are often masked one by the other. In the weekly figures of the Bank of England, Jevons discovered monthly, quarterly, and annual periods.*

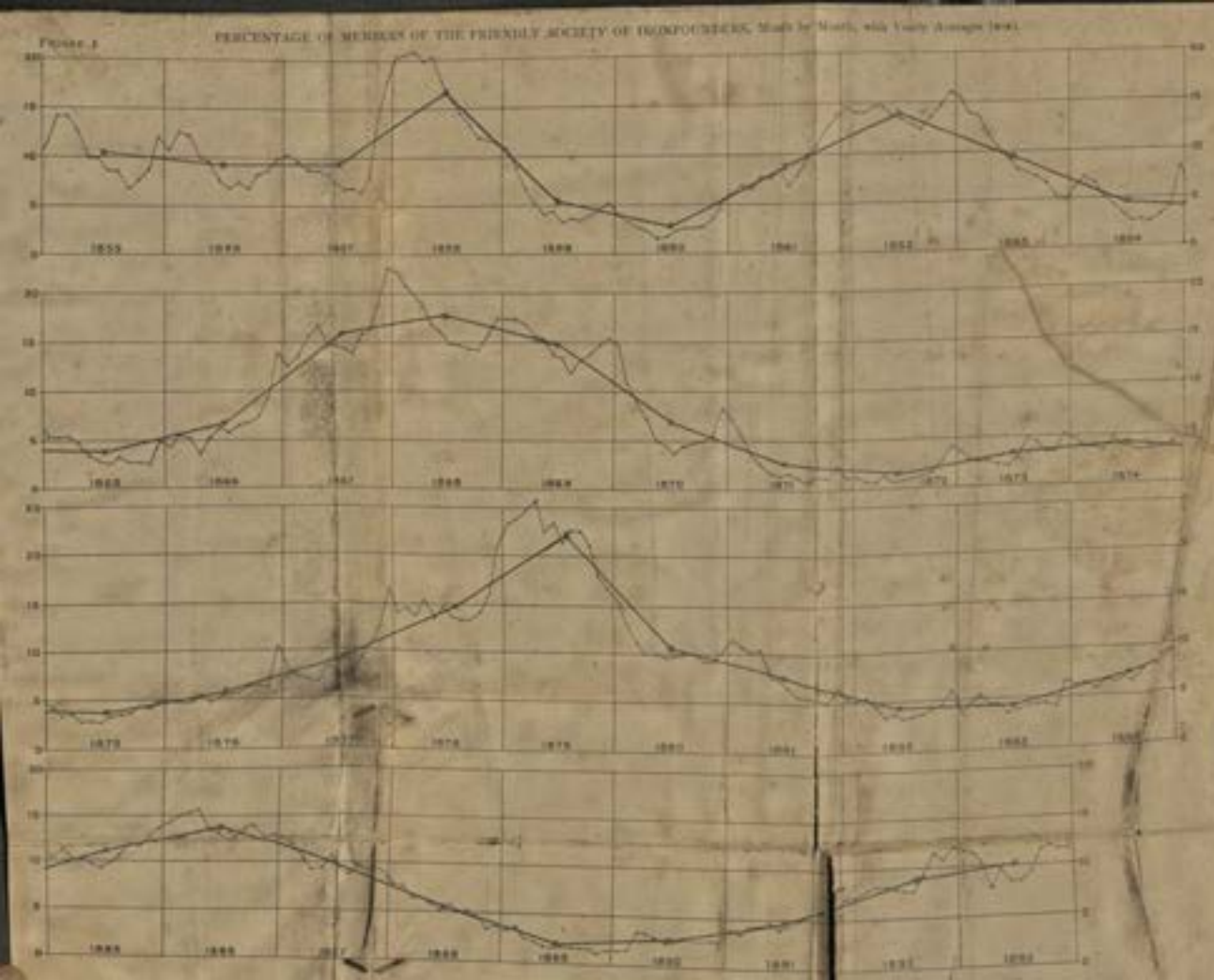
Periodic figures.

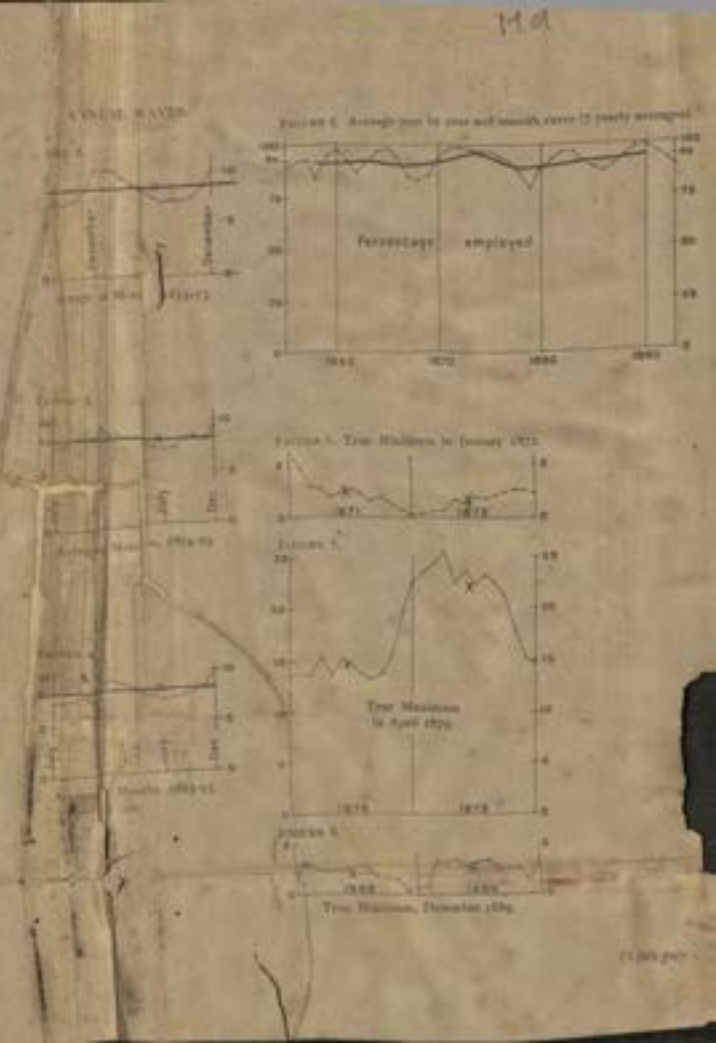
In social and industrial statistics we usually find an annual period, combined with a general slow movement upwards or downwards, and confused by an irregular period of about ten years, due to alternate inflation and depression of trade. The influences of these three movements on the resulting numbers can be investigated, and the general methods of examining periodic figures fully explained by the complete discussion of one example, viz., the monthly returns of want of employment of the Friendly Society of Ironfounders. For another example the reader is referred to Jevons' essay, *On the Frequent Autumnal Pressure in the Money Market;** and for an exercise, to the monthly gazette wheat prices, where the gradual change of the shape of the annual diagram can be traced in relation with the increasing influence of harvests in all the quarters of the globe.

These figures are specially suitable for showing graphically a double period, and the influences of rapid annual fluctuations and general movements of longer period on each other. Looking at the table on p. 179 along the lines for the several years, it will be seen that there is always a fall in the middle of the year. Looking down a vertical column under any

General features of the figures.

* See *Investigations in Currency and Finance.*


PERCENTAGE OF MEMBERS OF THE FRIENDLY SOCIETY OF IRONFOUNDERS, Made by Month, with Yearly Averages [etc].


Percentage employed

NUMBER OF UNEMPLOYED IRONFOUNDERS, expressed as percentages of estimated total number of members, month by month: calculated from figures given in the Annual Report of the Friendly Society of Ironfounders, 1894.

Year.	Jan.	Feb.	Mar.	April.	May.	June.	July.	Aug.	Sept.	Oct.	Nov.	Dec.	Average for Year.
1855	11.1	14.1	14.0	12.5	10.0	9.9	8.7	8.7	6.8	7.7	8.8	12.0	10.4
1856	10.9	12.6	12.2	10.0	9.4	7.5	6.9	7.3	6.9	8.1	8.7	9.9	9.2
1857	10.1	9.5	8.7	8.7	8.1	7.3	6.8	6.9	6.2	8.0	14.0	17.7	9.3
1858	20.2	20.6	20.9	19.8	20.3	17.8	15.9	14.3	13.1	11.9	11.5	11.2	16.5
1859	10.6	8.8	6.5	5.2	4.0	4.4	3.2	3.6	3.4	3.8	4.6	5.1	5.3
1860	4.0	3.2	2.6	2.2	1.6	1.7	2.3	2.6	2.6	2.9	3.7	5.6	2.9
1861	6.0	6.9	6.5	7.9	7.8	8.4	6.9	7.9	9.5	10.7	12.4	13.8	8.7
1862	14.5	14.0	14.0	14.6	14.4	13.7	13.3	12.9	12.2	13.5	14.9	16.0	14.0
1863	15.5	13.9	13.6	11.6	10.4	9.3	8.1	7.8	7.4	6.6	5.3	5.0	9.5
1864	6.0	7.1	6.6	5.1	4.4	3.3	2.8	2.8	2.6	3.3	4.2	8.1	4.7
1865	5.4	5.3	5.3	4.6	3.4	2.9	2.6	3.1	2.7	2.6	2.3	4.9	3.8
1866	4.2	5.4	5.1	3.6	5.1	6.5	5.9	6.5	6.9	7.4	9.3	13.8	6.7
1867	12.4	13.2	15.4	16.7	14.9	14.6	14.2	13.9	15.7	16.3	18.9	22.6	15.7
1868	22.1	20.9	19.8	18.6	16.7	15.8	14.9	14.7	14.2	14.1	15.6	17.4	17.1
1869	17.3	17.1	16.8	15.6	15.2	13.6	13.3	11.8	13.1	13.6	14.8	15.3	14.8
1870	14.5	10.9	8.7	7.2	5.0	4.5	3.7	4.5	4.9	5.0	5.6	8.3	6.9
1871	7.2	5.6	3.6	2.8	1.6	1.5	1.6	1.2	.9	1.4	1.1	2.2	2.6
1872	1.1	1.1	.9	.8	1.2	.7	.9	1.0	1.3	1.8	2.6	4.1	1.5
1873	3.3	2.8	2.7	2.5	2.1	2.0	3.0	4.9	4.3	3.3	3.3	5.1	3.3
Average 1855-73	10.3	10.2	9.7	8.9	8.2	7.7	7.1	7.2	7.1	7.5	8.5	10.4	8.6
1874	4.9	3.9	3.9	3.5	4.9	3.9	3.8	3.4	3.5	3.7	3.9	5.0	4.0
1875	4.6	3.4	3.5	2.8	2.8	2.8	3.3	3.4	3.6	4.1	4.1	5.0	3.6
1876	4.9	4.9	4.9	5.4	4.8	5.2	5.7	5.8	6.4	6.4	6.2	10.3	5.9
1877	7.7	7.4	7.0	6.9	8.4	7.6	7.4	7.8	9.6	10.9	12.3	16.3	9.1
1878	14.0	14.3	13.5	15.3	13.3	14.6	13.6	13.2	13.3	14.0	15.7	21.0	14.7
1879	23.2	23.8	24.7	25.5	22.3	23.4	21.5	22.6	22.5	21.1	18.0	16.6	22.1
1880	15.2	12.9	11.1	10.0	10.0	9.7	9.8	10.0	10.0	9.2	9.2	10.2	10.6
1881	11.5	10.8	10.1	10.1	7.6	7.5	6.5	5.8	5.6	5.4	5.0	6.6	7.7
1882	5.5	5.2	5.3	4.5	3.6	3.8	3.2	3.4	3.6	4.1	4.4	6.0	4.4
1883	3.6	4.8	5.2	4.3	4.2	3.6	3.9	4.3	4.3	4.2	4.0	6.6	4.4
1884	6.1	6.2	5.9	6.5	6.5	6.9	6.5	7.6	8.1	7.8	9.8	10.9	7.4
1885	10.2	11.1	10.0	10.1	9.8	9.1	9.8	10.7	11.8	11.6	12.7	13.6	10.9
1886	14.1	15.0	15.2	15.5	13.4	13.1	12.1	12.7	13.6	13.9	12.7	12.9	13.7
1887	12.4	11.6	10.2	9.1	9.2	10.6	9.2	8.8	9.6	9.4	9.4	9.1	9.9
1888	7.8	7.5	6.4	6.4	5.9	5.2	5.7	5.0	5.1	4.8	3.2	3.5	5.5
1889	3.1	3.3	2.4	2.1	1.7	1.6	1.7	1.7	1.6	1.5	1.2	1.4	1.9
1890	1.3	1.3	3.2	3.1	2.8	2.4	2.4	2.7	2.7	2.7	2.7	2.7	2.5
1891	3.9	3.5	4.2	4.2	4.6	4.0	4.5	4.8	5.4	5.6	5.7	6.3	4.7
1892	7.0	7.2	7.9	8.1	7.9	7.9	7.7	7.6	9.3	11.4	10.9	12.0	8.7
1893	11.5	11.2	10.1	7.7	9.6	8.3	8.3	9.2	11.7	11.9	11.5	11.5	10.2
Average 1874-93	8.6	8.5	8.2	8.1	7.7	7.6	7.3	7.5	8.1	8.2	8.1	9.4	8.1
Average 1855-93	9.4	9.3	8.9	8.5	7.9	7.6	7.2	7.4	7.6	7.9	8.3	9.9	8.3

month, it will be seen that there is no generally marked tendency towards increase or diminution, for high and low numbers occur in the first as well as the last few years. The most noticeable feature of these figures is the alternation of groups of years of high and of low numbers. Percentages above 10 will be found in 1861-63, 1866-70, 1877-81, 1884-87, and 1892-93. Let us choose for examination the period 1866-70. The figure for January 1866 is below the Januaries of previous years; those of February, March, and April are also low; from May to September the figures are greater than those of 1865 or 1864; from October to December they are greater than those of 1863, 1864, or 1865; in December 1867 they are greater than any previous year. Most of the figures for 1868 beat the record up to that date; but from September 1868 the figure is lower than the one twelve months earlier till July 1872. This wave of unemployment then lasted from May 1866 to September 1872.

Seasonal influence.

Now let us watch the seasonal influence. In 1866 there was no fall in the summer except in April, and there was a very rapid rise in December. In 1867 a fall in May and a slight fall from June to August was followed by a rapid rise in November and December. There is a fall from December 1867 to September 1868, but a rise follows in October, November, and December; since the rise does not generally begin till August, it will be seen that the general fall did not much delay the seasonal effect. In the next year, 1869, there is a fall to a lower minimum in August, but now the rise in December is very slight, next year the fall is very quick to August, but the seasonal rise is not delayed. From this it is clear that the seasons had their effect throughout the fluctuation except in the opening year 1866, when there was no fall, and that the rises in the autumn were very much accentuated. Almost identical remarks would apply to the period August 1875 to May 1881. In what month was the condition of employment 1867-70 at its worst? The greatest figure given is 22.6 per cent. in December 1867, but unemployment in December is generally greater than in any other month, and the figures for any of the following six months may be more unusual; the determination of the exact date will be best shown by diagrams. It may be mentioned that most of these remarks were suggested by Mr Hey, the former secretary of the Iron-founders' Society, who drew up these figures.

If we now turn to the diagram, the following facts may be noticed. The thick line showing the annual average percentages shows a downward tendency till 1857, followed by an abrupt rise and fall in 1858, then three years' rise to its original height, returning to a minimum in 1865; the next wave covers six years, and is marked by an extraordinarily sharp rise in 1867, and a very low minimum in 1872. The exceptional condition of trade in 1872 could not last, but the rise is very gradual to 1876, when the next cycle of trade is marked again by a six years' wave: the rise is not so steep as in the former fluctuation, but lasts longer, and a higher point is reached: the fall is at about the same angle, and the minimum in 1882 is about the same as that in 1865. The next wave came before it appeared to be due, and lasted seven instead of six years, but was much more moderate, and again the rise was sharper than the fall. The minimum of 1889 did not endure, and the figure ends with a suggestion that the maximum will be in 1894, but only at a moderate height, and the next minimum might be expected in 1898 or 1899, if causes similar to those which influenced earlier trade depressions were still acting. It may be found, in fact, from the Board of Trade returns, that, taking all the trade unions who made returns together, the maximum month was December 1892, and the maximum year was 1893; after this the fall is regular to 1897, and a trifling rise in 1898 is followed by a very low figure for 1899.*

The story from the diagram.

In figure 5 the diagram is inverted and greatly compressed, showing now the percentage employed. If the period 1876-82 is cut off by two vertical lines, readers may see how great were the amounts of labour lost to the country and wages to the members of the Ironfounders' Society in those years. These figures show a want of employment due to special causes in this Society more than twice as great as in other Unions whose returns are available for the same period.

In figure 5 the annual averages are smoothed by the method explained above (p. 152), a seven-yearly average† being taken

* See *Annual Abstract of Labour Statistics*, 1895, p. 73, for various methods of treating these figures similar to those here discussed.

† For smoothing and studying periodic curves, see Professor Poynting's paper in *Statistical Journal*, 1884.

to correspond to the general wave length. It will be seen that there is no very marked tendency up or down in the thirty-nine years, and that the smooth line is never far from the general average of employment, 91.7.

The comparison of this diagram with that illustrating exports (p. 151) is very instructive. Some of the results may be thus exhibited:—

Dates of		Dates of	
Minima of Exports.	Maxima of Unemployment.	Maxima of Exports.	Minima of Unemployment.
1862	1858 and 1862	1866	1865
1868	1868	1872	1872
1879	1879	1882	1882 or 1883
1886	1886	1890	1889
1894	1893		

The figures may also be compared graphically by the methods of the previous or following sections.

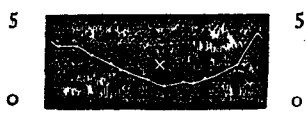
Measurement of seasonal influence.

The averages for the nineteen Januaries, nineteen Februaries, &c., in the years 1865-73, and similar averages for the years 1874-93, and the whole period are given in the table and exhibited in figures 2, 3, 4. When we calculated the annual averages just discussed we eliminated by that process the seasonal fluctuations; by this new series of averages we eliminate the influences of particular years. If we took, for instance, all the November numbers out of a series of figures totally uninfluenced by the seasons, if such could be found, and compared these with the general average for all months, we should in the long run find just as many instances above as below this average; but if the figures were influenced by the seasons, we should find a considerably greater number above than below, or *vice versa*. The greater the seasonal influence, the greater would be this excess or defect. Averaging numbers in this way eliminates the non-seasonal causes, for by hypothesis the excesses and defects due to them will in the long run balance one another; and except by averaging these cannot be eliminated, unless they can be actually calculated. The excess of the November average above the general average will be greater than that of October, if the seasonal causes exert more influence towards excess in the former than in the latter month, and the curve which shows these averages will show a resemblance to that which would

be obtained if the non-seasonal causes were absent. It will be only a resemblance for two reasons: first, because in the comparatively short series of years with which we are generally obliged to be content, a very effective non-seasonal cause will leave its mark on the average, as may be seen in the table on p. 179; secondly, because seasonal and non-seasonal causes are often not independent; a depression of trade is accentuated by a sharp winter; a bad season in a year of bad trade may increase the want of employment greatly and suddenly, while a good summer in a prosperous year may reduce it almost to zero. In the case we are considering the interaction of causes tends to exaggerate the seasonal maximum and diminish the minimum; in other cases a contrary effect might be found.

In figures 2, 3, 4 the curve for the latter half of the year is prefixed to that of the calendar year, because the character of the yearly waves is seen most clearly from minimum to minimum. It may be noticed that the wave in figure 3 is less definite in shape and has a smaller rise and fall than that of the earlier period shown in figure 2; it would appear that the seasons are losing their influence.

If there is a definite annual period, that represented by figure 4, it may be expected that a figure of a shape similar to this—



will be repeated annually in figure 1; it is shown well in 1864, 1882, and other years. In the great majority of cases the yearly maximum is reached in December or January; at the end of 1858 the maximum is absent, but is replaced by a break in the rapidity of the fall; at the end of 1860 there is a rise, but the spring fall following is checked by the general upward trend; similar remarks apply to all the great fluctuations. There is no doubt that right along the line we find at nearly equal intervals these pointed crests above the line of averages.

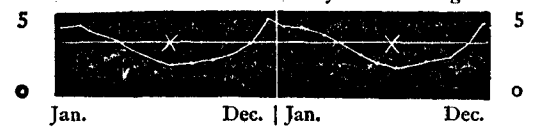
The annual wave.

The minima are not so conspicuous, for the pointed shape is absent, trifling causes bring them near the smoothed line, and they are easily masked by a general fall or are absent because of a general rise. In 1861, however, there is a distinct minimum

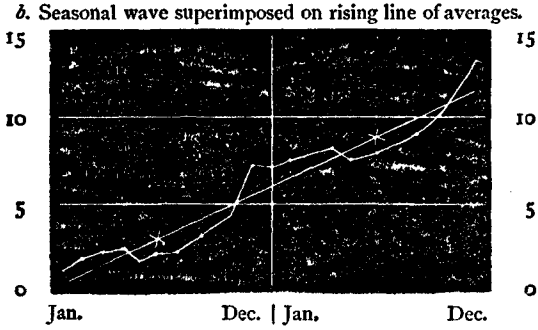
in spite of the strong upward tendency; the minima are very conspicuous throughout the fluctuation of 1865-70; and from 1859 to 1888 the minima are fairly marked, except in 1876, 1880, and 1881.

The following figures show the effect of a stationary, rising, and falling average annual rate on the shape of the seasonal wave:—

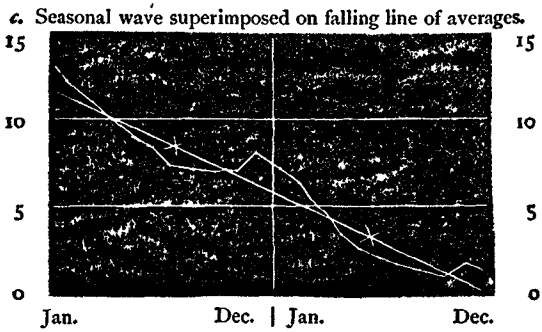
a. Seasonal wave on stationary line of averages.



b. Seasonal wave superimposed on rising line of averages.



c. Seasonal wave superimposed on falling line of averages.



These figures are drawn by adding or subtracting the average monthly differences from the general average.

viz.	Jan.	Feb.	Mar.	Apr.	May.	June.	July.	Aug.	Sept.	Oct.	Nov.	Dec.
	+1.1	+1.0	+.6	+.2	−.4	−.7	−1.1	−.9	−.7	−.4	0	+1.6

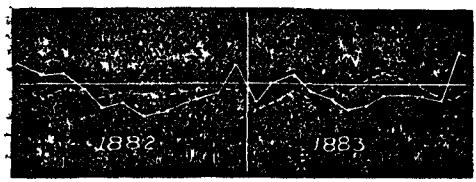
month by month to or from the positions shown on the straight lines joining the annual averages. On a rising line the spring

fall tends to become horizontal and the autumn rise steeper; on a falling line the spring fall becomes more rapid and the autumn rise is checked.

If this seasonal wave, added to the slower long-period changes, were the complete explanation of these numbers, figure 1 (p. 179) would be entirely composed of modifications of figures *a*, *b*, and *c*. Figure *a* is exemplified especially in 1855-57, 1864-65, 1871-73; figure *b* in 1860-61, 1866-67, 1877-78, 1883-85; figure *c* in 1859, 1863, 1880-82, 1886-89.

As explained above, the two sets of causes are not independent, and these figures are not reproduced exactly; but the resemblance is sufficiently close to make the following method of eliminating seasonal fluctuations partially applicable. Combine the monthly excesses and defects just given with the original numbers, by subtracting the excesses and adding the defects; this process should tend to produce a straight line thus:—

Elimination of fluctuations.

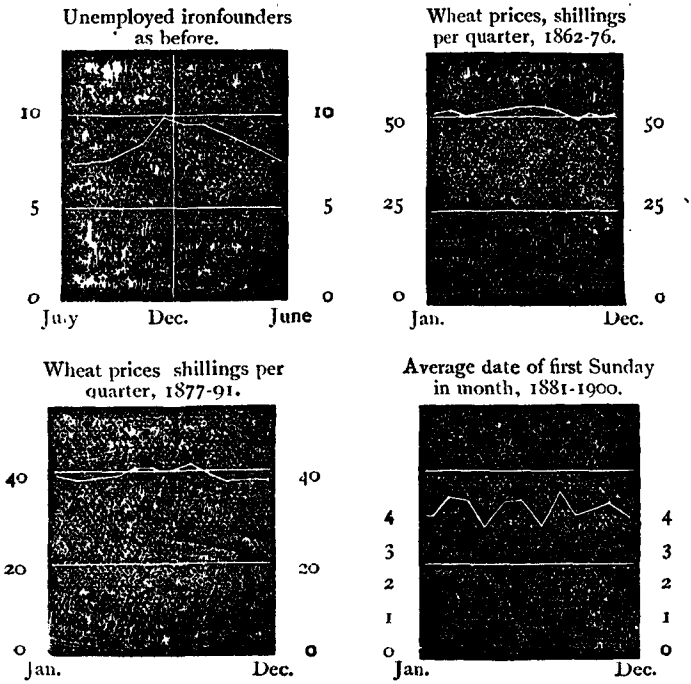


But the result is not more than a tendency, because of the unusual fall in January 1883, and it is difficult to find a perfect example. This method is applied in figures 6, 7, and 8 in an attempt to disentangle the seasonal fluctuations from the effects of the commercial crisis of 1872, the depression of 1879, and the turn of the tide in 1883. In figure 6 it is seen that January 1872 was the best month relatively, though the absolute minimum was not reached till June of that year; from this it appears that January 1872 was the turning point of the great inflation, a date somewhat earlier than that generally given. The date of the maximum of 1879 is left unchanged by this process, and that of the 1889 minimum is only shifted one month.

We have still to discuss the criteria of the existence of a period. In figure 1 the optical evidence is sufficient to suggest the annual period, but it may be doubted whether an annual fluctuation would be suggested by a diagram representing wheat prices. It is clear that if the

Criteria of existence of period.

monthly entries of any returns whatever were averaged in months over any period of years, that the averages for January, February, &c., would not be exactly equal, even if there were no seasonal influence. The following diagrams show various averages:—



Of these the first three may be expected to be seasonal, while the last, which shows the averages of the dates on which fell the first Sunday in 20 Januaries, 20 Februaries, &c., in a series of years, certainly is not.

The following simple tests may be applied to decide this point. If the period is in any way connected with the seasons, it will correspond to some extent to the ordinary weather charts of temperature, &c., which have a single annual maximum and corresponding minimum. Phenomena affected by the weather may also be expected to show a single maximum, nearly coinciding with the maximum or minimum temperature; thus the maximum unemployed coincides with the minimum length of daylight and precedes the minimum temperature. In some

cases a second subsidiary maximum may be shown, since, for example, an excessive death rate may be due to excessive cold or heat; but even in this example further analysis would probably show that the one maximum was for the old, the other for the young. Wheat prices may also show two minima due to the harvests in the two hemispheres. The "Sunday" curve just given shows four maxima, and is not seasonal. More than one maximum is evidence against periodicity till their existence can be explained.

The second test is to look at the serial diagram and notice how often the maximum occurs in the same month; non-periodic causes will hide the maximum occasionally, but in the long run one month will be predominant. In figure 1 the maximum occurs in March and April twice each, in February three times, in January eleven times, and in December twenty-one times. The maximum is then generally in midwinter. The minimum is not in this case so well defined.

Probability test.

The following table shows how this analysis can be extended:—

	Times out of 39.
The percentage of December is greater than that of the preceding November - - - - -	33
The percentage of December is greater than that of the following January - - - - -	28
The percentage of December is greater than that of the preceding July - - - - - -	33
The percentage of December is greater than that of the following July - - - - - -	30

The chances against so great a preponderance, if the seasons had no influence, are respectively 70,000 to 1, 106 to 1, 70,000 to 1, and 940 to 1.* All the months may be separately tested in the same way. This method by no means exhausts the evidence, for we have only considered which of two months is the greater, and not how great is the excess when it exists. On this point the reader is referred to the paper by Professor Edgeworth, *On Methods of Statistics*, in the Jubilee Volume of the Royal Statistical Society, p. 206; this should, however, be postponed till the mathematical treatment which follows in Part II. has been studied.

* See Part II., Sect. I., *infra*.

5. Logarithmic Curves.

Need for graphic representation of ratios.

A serious flaw in the graphic method as used in the previous sections is that, when we are dealing with a series of increasing figures, though the totals year by year may be increasing, we are compelled to represent equal increments on these totals by equal vertical distances; thus an increment of £20 on a total of £20 is represented by the same vertical distance as an increment of £20 on a total of £2,000. Thus in the annexed figure representing exports, the fall from £52,000,000 to £42,000,000 in 1815-16 is barely noticeable, though it is a fall of 20 per cent., and was connected with very great distress in the manufacturing districts; while the fall from £305,000,000 in 1883 to £269,000,000 in 1886 attracts attention immediately, though it is one of 12 per cent. only. Again the increase of 34 per cent. which took place between 1848 and 1850 appears insignificant in comparison with that of 29 per cent. from 1870 to 1872. When we are attacking questions of causation it very frequently happens that we are more concerned to know the proportionate increase than the actual increase. When we are considering the gradual growth of our foreign trade, or when we are comparing the growth of trade of two countries, a diagram like that annexed is likely to give quite a wrong impression of the struggle that marked the early stages. We need then a diagram not of quantities, but of ratios, where equal vertical distances represent no longer equal absolute increments, but equal proportional increments, that is, equal rates of increase. By the use of logarithms a universal scale can be constructed which serves this purpose. The non-mathematical student can easily accustom himself to the use of diagrams so constructed, by studying one where the actual amounts represented are entered, and noticing that whatever part of the scale he takes, doubling, halving, increasing by 20 per cent. and so on, are always represented by the same vertical distances respectively. The construction of a diagram on this scale is as follows:—Write down the numbers in the series to be represented; against them write down their logarithms; on paper divided into equal squares mark at equal intervals on a

Construction of a logarithmic diagram.

vertical line numbers ascending in regular progression so as to include all the logarithms found; mark off the dates on a horizontal line; and on the scale thus prepared mark in the logarithms, instead of the original numbers. The table on p. 191 and the diagram facing p. 190 show the figures of imports and exports thus treated. On the right hand of fig. 2 the position of the absolute numbers is given; on the left the corresponding logarithms. A given vertical distance, 1 inch, represents the distance .301 on the logarithmic scale; if we add this quantity to the logarithm of any number, we obtain the logarithm of twice that number for $\log a + .301 = \log a + \log 2 = \log 2a$; for instance, if we increase the height of the position which represents £30 by 1 inch, we arrive at the position which represents £60. Again if we now add 1.59 of an inch, which represents .477 on the same scale as before, that is log 3, to the logarithm of $2a$, we obtain log $6a$, and we have—

$$\log 6a = .477 + \log 2a = .477 + .301 + \log a, \text{ as above}$$
$$= .778 + \log a = \log 6 + \log a;$$

that is, we arrive at the same position on this scale whether we go by means of two separate ratios or by a single compounded ratio. Thus a diagram drawn on this principle satisfies the necessary conditions that equal vertical distances represent the same ratio in whatever part of the scale they are taken, and that any number of points can be entered without leading to inconsistencies. At the end of this section is given a table of the logarithms of 1 to 1,000, correct to the third decimal place, which will be found sufficient for this purpose.

Thus on the diagram given we can see at once that imports were doubled in value between 1810 and 1836, again between 1840 and 1853, again between 1855 and 1866, and that their value increased 40 per cent. between 1886 and 1899. Or we may notice that the excess of the value of imports over that of exports was 40 per cent. of the latter both in 1850 and in 1880; that the value of imports in 1899 was thrice that of exports in 1860.

Examples of its use.

If the eye has been carefully educated to understand a diagram of this sort, if the fact that it is a *diagram of ratios, not of quantities*, is firmly impressed on the mind, then the diagram answers perfectly the object of the graphic method, that is, it gives a true instantaneous impression of a complex series of facts. If, on the other hand, it is found that a true

impression is not received, through inability to take the right mental position, then diagrams on the natural scale should be employed only, always with the recollection that they may give false impressions of ratio.*

Velocity and acceleration.

It is to be noticed that no base line should be given in diagrams of this class, otherwise a false impression is at once obtained. Notice further that, while equal vertical differences represent equal ratios from any part of the diagram to any other, instead of equal increments as on the natural scale, equal degrees of slope represent equal ratios of increase (equal accelerations), instead of equal additions in equal times as on the natural scale (equal velocities). On the logarithmic scale a line rising with convexity to the horizontal shows that the ratio of increase is growing, as in imports from 1830-1853 (if the line is smoothed), while concavity, as from 1854 to 1873, shows a slackening; but on the natural scale the line is convex almost throughout the two periods, showing that the actual increments were increasing all the time.

Useful application to index-numbers.

It would be useful, if space permitted, to offer several diagrams on both scales; for in many series of figures the differences exhibited by the two methods are very instructive. One case may be signalized where the logarithmic scale is specially important, that is, when the original numbers represent ratios, not actual numbers. Thus in Mr Sauerbeck's well-known diagram, drawn on the natural scale, representing his index-numbers of prices, all the numbers included are percentages of their values in certain defined years. Suppose that 100, 80, and 60 are the index-numbers for three years, then on the natural scale the decrements are represented by equal distances and appear to be equal. The changes in the value of gold, however, are by no means equal in the two periods. In the first, the fall from 100 to 80 is one of 20 per cent.; 16s. at the second date would buy goods which cost £1 at the first. In the second, the fall from 80 to 60 is one of 25 per cent.; 15s. at the last date would buy goods which cost £1 at the middle date. For the purposes of price index-numbers it is ratios which are important and which the diagram should represent.

* Professor Marshall suggests a simple method of correcting this false impression in his paper *On the Graphic Method of Statistics*, in the jubilee volume of the *Journal of the Royal Statistical Society*, p. 257 *seq.*

FIGURE I.

GROWTH OF IMPORTS AND EXPORTS.
I. ON NATURAL SCALE.

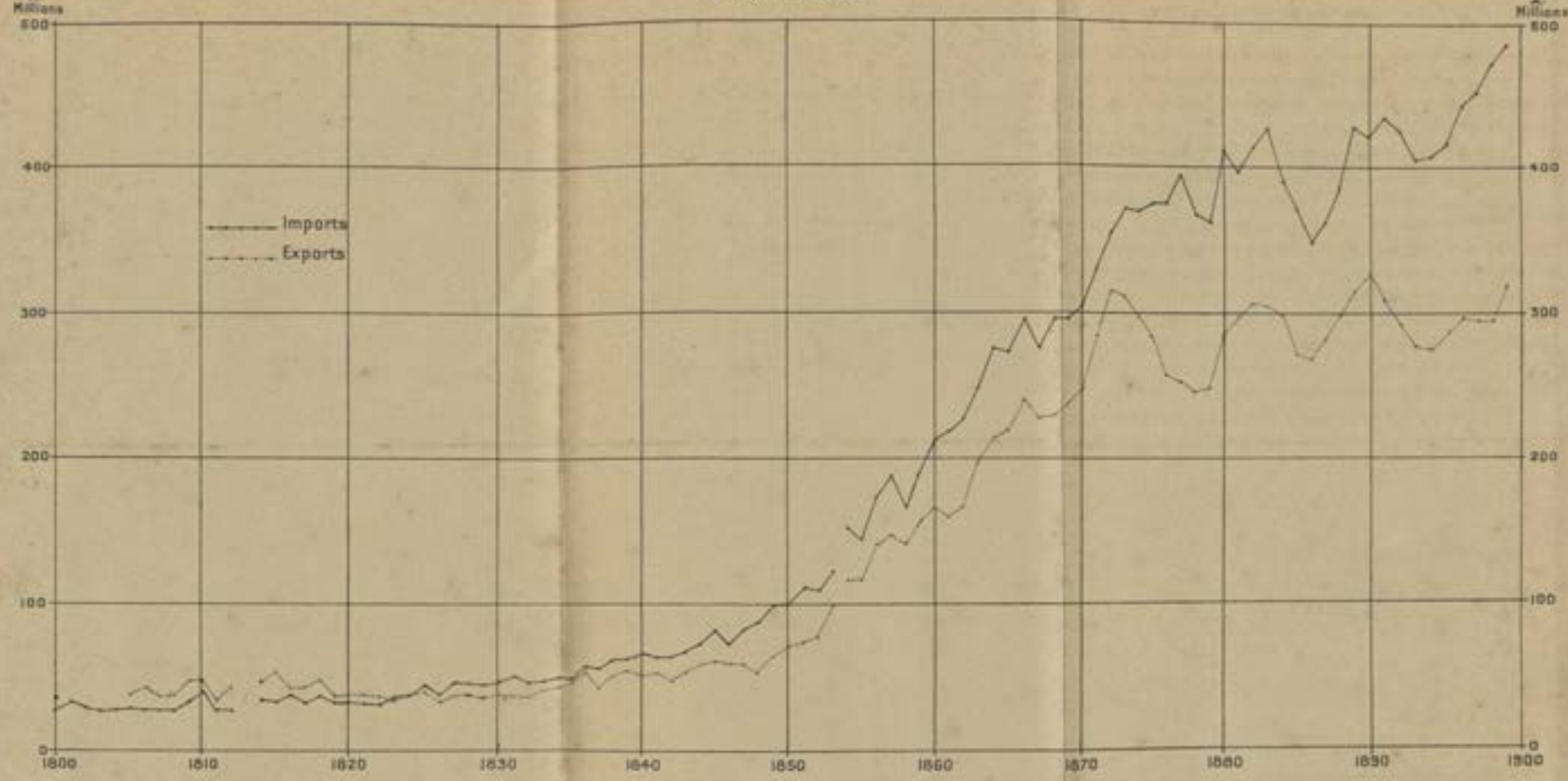
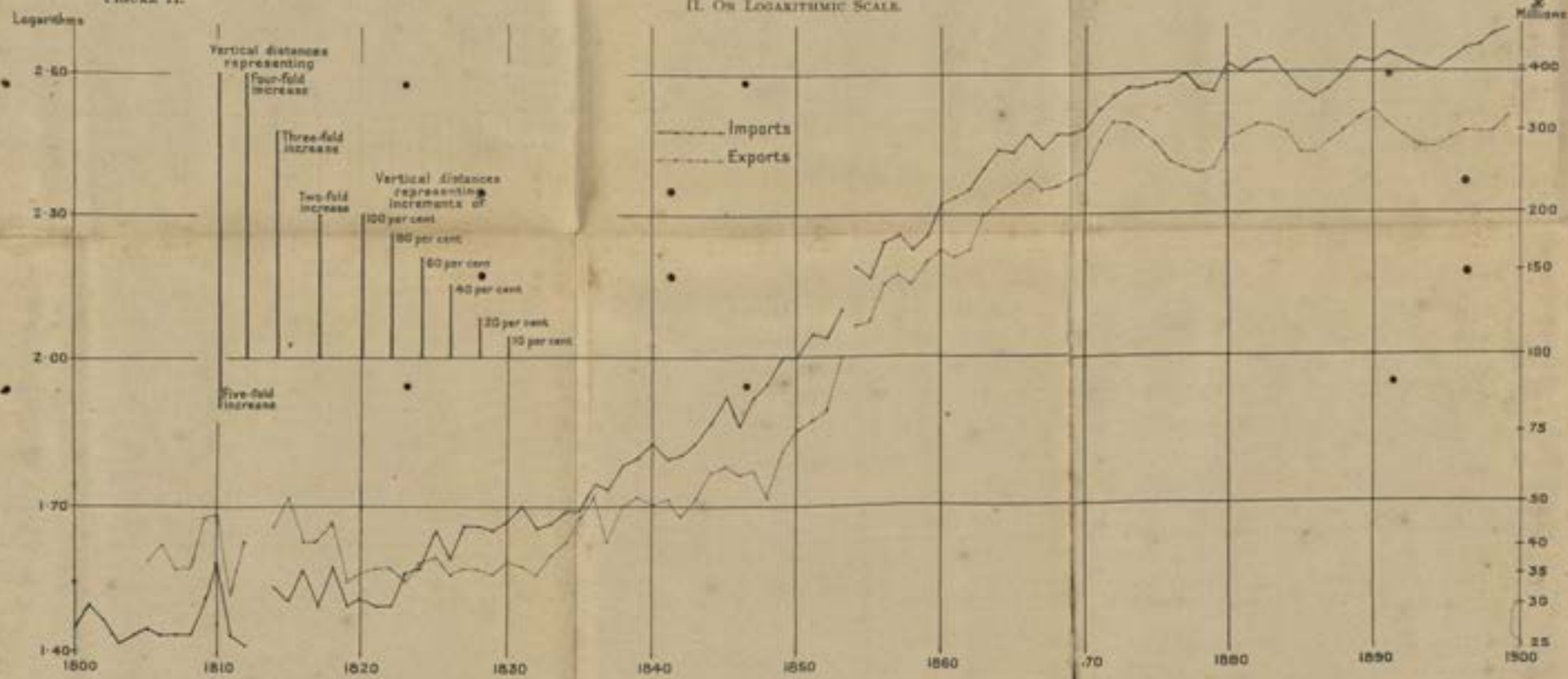


FIGURE II.

II. ON LOGARITHMIC SCALE.



To face page 190.

Year.	Imports.*	Logarithms.	Exports.†	Logarithms.	Year.	Imports.*	Logarithms.	Exports.†	Logarithms.
1800	28	1.447	34	1.531	1850	100	2.000	71	1.851
1801	31	1.491	...	...	1851	111	2.045	74	1.869
1802	29	1.462	...	...	1852	109	2.037	78	1.892
1803	26	1.415	...	...	1853	123	2.090	99	1.996
1804	27	1.431	...	...	1854	152	2.182	116	2.064
1805	28	1.447	38	1.580	1855	144	2.158	117	2.068
1806	27	1.431	41	1.613	1856	173	2.237	139	2.143
1807	27	1.431	37	1.568	1857	179	2.252	146	2.164
1808	27	1.431	37	1.568	1858	165	2.216	140	2.146
1809	32	1.505	47	1.672	1859	179	2.252	156	2.193
1810	39	1.591	48	1.681	1860	211	2.324	165	2.217
1811	27	1.431	33	1.518	1861	217	2.336	160	2.204
1812	26	1.415	42	1.623	1862	226	2.354	166	2.220
1813	...	...	...	...	1863	249	2.396	197	2.295
1814	34	1.531	45	1.653	1864	275	2.439	213	2.328
1815	32	1.505	52	1.716	1865	271	2.432	219	2.340
1816	37	1.568	42	1.623	1866	295	2.470	239	2.379
1817	31	1.491	42	1.623	1867	275	2.439	226	2.354
1818	37	1.568	46	1.663	1868	295	2.470	228	2.358
1819	31	1.491	35	1.544	1869	295	2.470	237	2.375
1820	32	1.505	36	1.556	1870	303	2.481	244	2.387
1821	31	1.491	37	1.568	1871	331	2.519	284	2.453
1822	31	1.491	37	1.568	1872	355	2.550	315	2.498
1823	36	1.556	35	1.544	1873	371	2.569	311	2.492
1824	37	1.568	38	1.580	1874	370	2.568	298	2.475
1825	44	1.643	39	1.591	1875	374	2.573	282	2.450
1826	38	1.580	32	1.505	1876	375	2.574	257	2.410
1827	45	1.653	37	1.568	1877	394	2.596	252	2.401
1828	45	1.653	37	1.568	1878	369	2.567	245	2.389
1829	44	1.643	36	1.556	1879	363	2.559	249	2.396
1830	46	1.663	38	1.580	1880	411	2.614	286	2.456
1831	50	1.699	37	1.568	1881	397	2.599	297	2.473
1832	45	1.653	36	1.556	1882	413	2.616	307	2.487
1833	46	1.663	40	1.602	1883	427	2.630	305	2.484
1834	49	1.690	42	1.623	1884	390	2.591	296	2.471
1835	49	1.690	47	1.672	1885	371	2.569	271	2.432
1836	57	1.756	53	1.724	1886	350	2.544	269	2.429
1837	55	1.740	42	1.623	1887	362	2.558	281	2.448
1838	61	1.785	50	1.699	1888	388	2.586	299	2.476
1839	62	1.792	53	1.724	1889	428	2.631	316	2.500
1840	67	1.825	51	1.708	1890	421	2.624	328	2.516
1841	63	1.799	52	1.716	1891	435	2.638	309	2.490
1842	64	1.806	47	1.672	1892	424	2.627	292	2.465
1843	68	1.832	52	1.716	1893	405	2.607	277	2.442
1844	74	1.869	59	1.771	1894	401	2.603	274	2.439
1845	83	1.919	60	1.778	1895	417	2.620	286	2.456
1846	73	1.863	58	1.763	1896	442	2.645	296	2.471
1847	83	1.919	59	1.771	1897	451	2.654	294	2.468
1848	89	1.949	53	1.724	1898	471	2.673	294	2.468
1849	100	2.000	64	1.806	1899	485	2.685	320	2.505

* Imports—Official values till 1853; real values from 1854.

† Including re-exports.

The logarithmic scale has special uses in the comparison of series of figures, and the methods discussed in the section devoted to that subject can be readily adapted.

Comparisons on the logarithmic scale.

The difficulty of the choice of units in comparing quantities of different natures disappears when we deal only with ratios; we need no longer trouble about the method of percentages. In investigating causal relations we are more likely to find close connection in ratios than in quantities; for if one set of phenomena is connected with another, it is more likely that the relation will be a proportional one (*e.g.*, that an increase of 10 per cent. in some measurable characteristic of the one corresponds to an increase of 8 per cent. in a characteristic of the other), than an absolute quantitative one (*e.g.*, that an increase of 2s. in a price, at whatever point it stands, corresponds to a decrease of 100 in the number of purchasers). Resemblance between two curves on the logarithmic scale will mean the correspondence in proportional change, while resemblance on the natural scale means correspondence in absolute change.

There is less trouble in this new method in equating averages than before. For if the logarithms of two series are taken, it is quite immaterial at what height on a logarithmic scale the two are plotted out; alteration of height only means multiplication of all the items by a constant quantity, and does not alter the appearance or proportion of their fluctuations. The method to be employed is as follows:—Draw the curves representing two series of figures on a logarithmic scale; then shift the lower curve vertically upwards to and over the other, till the closest possible correspondence is obtained; draw it in in this position, and the two series can be accurately compared.

The following example employs this method with a further development, corresponding to that of p. 177, *supra*, where fluctuations are equated. In the earlier method we used the average as a position from which to measure the various items, and adapted the scales; a similar method might again be used, but it is more convenient to keep to one logarithmic scale, and now we have no base line to consider. Calculate the fluctuations much as before, but express them as *percentages* of the adjacent maxima before taking their average. In the following example it is found that a fluctuation of 8.4 per cent. in the number employed, in those trade unions

Equation of fluctuations.

FIG. 1. Comparison in 1865-93.
ON NATURAL SCALE.

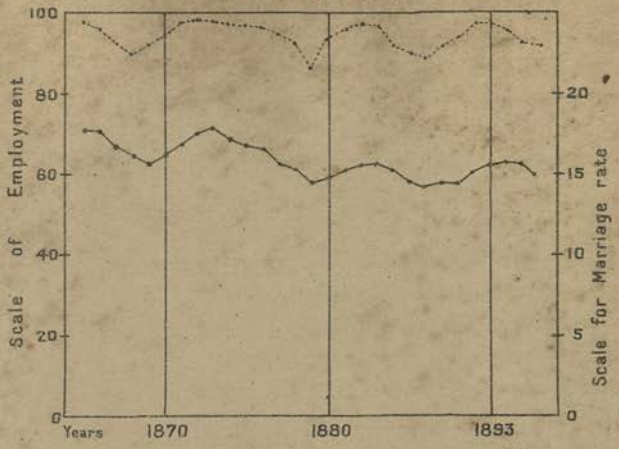


FIG. 2. The same; Logarithmic Scale.

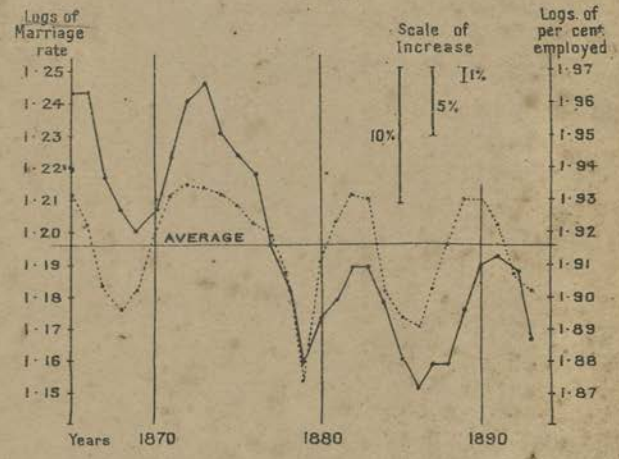
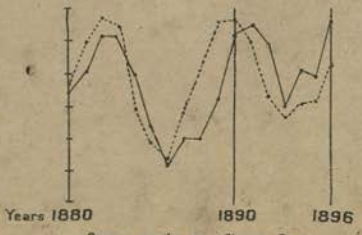


FIG. 3. Comparison in 1880-1896.



Same scale as figure 2.

Marriage rate ..... Employment

whose returns are accessible,* corresponds to one of 9.7 per cent. on the marriage rate. To investigate a possibly closer correspondence, assume that a portion of the number employed do not influence the marriage rate, and find what part must be subtracted before this 8.4 per cent. of the total forms as much as 9.7 per cent. of the remainder; the average percentage of members of the trade unions at work in the selected period was 95.1; 8.4 per cent. of this is 7.99, which forms 9.7 per cent. of 82.4. Thus 12.7, the difference between 95.1 and 82.4, may be considered as not influencing the question, and subtracted throughout before logarithms are taken. This process would be replaced on the natural scale by equating the averages of two series, and drawing one base line so far below the other that average fluctuations would be represented by the same vertical distance for both series; which process is exactly equivalent to that adopted on p. 177. Expressed algebraically, we are now investigating the equation—

$$\log (y - c) - \log x = k, \text{ a constant,}$$

where c and k are constants to be so selected as to give the closest fit, and y and x are the quantities to be compared.

In the adjacent diagrams, figure 1 gives the figures in the natural scale; figure 2 gives them on the logarithmic scale, after they have been arranged so as to make average percentage fluctuations equal; while in figure 3 the shorter period, 1880-96, is treated in a method precisely similar to that of figure 2. The actual numbers and logarithms are given on the next page.

Note.—In making comparisons of series, whether on the logarithmic or natural scale, instead of taking the average of the whole period and measuring deviations therefrom, it is legitimate and often advantageous to measure the deviations from a curve smoothed, as on p. 151, a line of moving averages. We are then ignoring the causes which have a gradual and permanent effect, and comparing the short-period fluctuations. See Mr R. Hooker's paper in the *Statistical Journal*, September 1901.

---

* The figures in columns 2 and 4 in the second table on the next page are taken from Mr G. H. Wood's paper on *Some Statistics of Working Class Progress since* 1860, *Statistical Journal*, 1900, where a valuable logarithmic diagram will be found, illustrating many of the points of this section.

Marriage Rate per 1,000.					Percentage Employed.				
Years.	Maxima.	Minima.	Differences.	% of Max.	Years.	Maxima.	Minima.	Differences.	% of Max.
1869	...	15.9	1.7	10	1868	...	91.5	7.4	7.5
1873	17.6	...	3.2	18	1872	98.9	...	11.4	11.5
1879	...	14.4	1.1	7	1879	...	87.5	10.6	10.8
1882-83	15.5	...	1.3	8	1882	98.1	...	7.6	7.8
1886	...	14.2	1.4	9	1886	...	90.5	7.4	7.6
1891	15.6	...	.9	6	1889-90	97.9	...	5.4	5.5
1893	...	14.7			1893	...	92.5		
				9.7					8.4

Average percentage employed, 1865-93, 95.1 ; 8.4 per cent. of 95.1 is 9.7 per cent. of 82.4.

Years.	Marriage Rate.	Logarithms.	Percentage Employed.	Less 12.7.	Logarithms.
1865	17.5	1.243	98.0	85.3	1.931
1866	17.5	1.243	96.9	84.1	1.925
1867	16.5	1.217	92.7	80.0	1.903
1868	16.1	1.207	91.5	78.8	1.896
1869	15.9	1.201	92.6	79.9	1.902
1870	16.1	1.207	95.7	83.0	1.919
1871	16.7	1.223	98.2	85.5	1.932
1872	17.4	1.240	98.9	86.2	1.935
1873	17.6	1.245	98.7	86.0	1.934
1874	17.0	1.230	98.2	85.5	1.932
1875	16.7	1.223	97.5	84.8	1.928
1876	16.5	1.217	96.4	83.7	1.923
1877	15.7	1.196	95.6	82.9	1.919
1878	15.2	1.182	93.7	81.0	1.908
1879	14.4	1.158	87.5	74.8	1.874
1880	14.9	1.173	94.1	81.4	1.911
1881	15.1	1.179	96.5	83.8	1.923
1882	15.5	1.190	98.1	85.4	1.931
1883	15.5	1.190	97.8	85.1	1.930
1884	15.1	1.179	92.6	79.9	1.902
1885	14.5	1.161	91.0	78.3	1.894
1886	14.2	1.152	90.45	77.7	1.890
1887	14.4	1.158	92.6	79.9	1.902
1888	14.4	1.158	95.2	82.5	1.916
1889	15.0	1.176	97.9	85.2	1.930
1890	15.5	1.190	97.9	85.2	1.930
1891	15.6	1.193	96.5	83.8	1.923
1892	15.4	1.187	93.7	81.0	1.908
1893	14.7	1.167	92.5	79.8	1.902
		Average 1.196			Average 1.916

Logarithms of Numbers 1 to 1,000, Correct to the Nearest Digit in the Third Decimal Place.

No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.
1	0	34	1.531	67	1.826	101	2.004	134	2.127	167	2.223	201	2.303	234	2.369	267	2.427	301	2.479
2	.301	35	1.544	68	1.833	102	2.009	135	2.130	168	2.225	202	2.305	235	2.371	268	2.428	302	2.480
3	.477	36	1.556	69	1.839	103	2.013	136	2.134	169	2.228	203	2.307	236	2.373	269	2.430	303	2.481
4	.602	37	1.568	70	1.845	104	2.017	137	2.137	170	2.230	204	2.310	237	2.375	270	2.431	304	2.483
5	.699	38	1.580	71	1.851	105	2.021	138	2.140	171	2.233	205	2.312	238	2.377	271	2.433	305	2.484
6	.778	39	1.591	72	1.857	106	2.025	139	2.143	172	2.236	206	2.314	239	2.378	272	2.435	306	2.486
7	.845	40	1.602	73	1.863	107	2.029	140	2.146	173	2.238	207	2.316	240	2.380	273	2.436	307	2.487
8	.903	41	1.613	74	1.869	108	2.033	141	2.149	174	2.241	208	2.318	241	2.382	274	2.438	308	2.489
9	.954	42	1.623	75	1.875	109	2.037	142	2.152	175	2.243	209	2.320	242	2.384	275	2.439	309	2.490
10	1.	43	1.633	76	1.881	110	2.041	143	2.155	176	2.246	210	2.322	243	2.386	276	2.441	310	2.491
11	1.041	44	1.643	77	1.886	111	2.045	144	2.158	177	2.248	211	2.324	244	2.387	277	2.442	311	2.493
12	1.079	45	1.653	78	1.892	112	2.049	145	2.161	178	2.250	212	2.326	245	2.389	278	2.444	312	2.494
13	1.114	46	1.663	79	1.898	113	2.053	146	2.164	179	2.253	213	2.328	246	2.391	279	2.446	313	2.495
14	1.146	47	1.672	80	1.903	114	2.057	147	2.167	180	2.255	214	2.330	247	2.393	280	2.447	314	2.497
15	1.176	48	1.681	81	1.908	115	2.061	148	2.170	181	2.258	215	2.332	248	2.394	281	2.449	315	2.498
16	1.204	49	1.690	82	1.914	116	2.064	149	2.173	182	2.260	216	2.334	249	2.396	282	2.450	316	2.500
17	1.230	50	1.699	83	1.919	117	2.068	150	2.176	183	2.262	217	2.336	250	2.398	283	2.452	317	2.501
18	1.255	51	1.708	84	1.924	118	2.072	151	2.179	184	2.265	218	2.338	251	2.400	284	2.453	318	2.502
19	1.279	52	1.716	85	1.929	119	2.076	152	2.182	185	2.267	219	2.340	252	2.401	285	2.455	319	2.504
20	1.301	53	1.724	86	1.934	120	2.079	153	2.185	186	2.270	220	2.342	253	2.403	286	2.456	320	2.505
21	1.322	54	1.732	87	1.940	121	2.083	154	2.188	187	2.272	221	2.344	254	2.405	287	2.458	321	2.507
22	1.342	55	1.740	88	1.944	122	2.086	155	2.190	188	2.274	222	2.346	255	2.407	288	2.459	322	2.508
23	1.362	56	1.748	89	1.949	123	2.090	156	2.193	189	2.276	223	2.348	256	2.408	289	2.461	323	2.509
24	1.380	57	1.756	90	1.954	124	2.093	157	2.196	190	2.279	224	2.350	257	2.410	290	2.462	324	2.511
25	1.398	58	1.763	91	1.959	125	2.097	158	2.199	191	2.281	225	2.352	258	2.412	291	2.464	325	2.512
26	1.415	59	1.771	92	1.964	126	2.100	159	2.201	192	2.283	226	2.354	259	2.413	292	2.465	326	2.513
27	1.431	60	1.778	93	1.968	127	2.104	160	2.204	193	2.286	227	2.356	260	2.415	293	2.467	327	2.515
28	1.447	61	1.785	94	1.973	128	2.107	161	2.207	194	2.288	228	2.358	261	2.417	294	2.468	328	2.516
29	1.462	62	1.792	95	1.978	129	2.111	162	2.210	195	2.290	229	2.360	262	2.418	295	2.470	329	2.517
30	1.477	63	1.799	96	1.982	130	2.114	163	2.212	196	2.292	230	2.362	263	2.420	296	2.471	330	2.519
31	1.491	64	1.806	97	1.987	131	2.117	164	2.215	197	2.294	231	2.364	264	2.422	297	2.473	331	2.520
32	1.505	65	1.813	98	1.991	132	2.121	165	2.217	198	2.297	232	2.365	265	2.423	298	2.474	332	2.521
33	1.519	66	1.820	99	1.996	133	2.124	166	2.220	199	2.299	233	2.367	266	2.425	299	2.476	333	2.522
				100	2.000					200	2.301					300	2.477		

LOGARITHMS OF NUMBERS 1 TO 1,000, CORRECT TO THE NEAREST DIGIT IN THE THIRD DECIMAL PLACE.

No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.
335	2.525	401	2.603	467	2.669	535	2.728	601	2.779	667	2.824	735	2.866	801	2.904	900	2.954
337	2.528	403	2.605	469	2.671	537	2.730	603	2.780	669	2.825	737	2.867	804	2.905	903	2.956
339	2.530	405	2.607	471	2.673	539	2.732	605	2.782	671	2.827	739	2.869	807	2.907	906	2.957
341	2.533	407	2.610	473	2.675	541	2.733	607	2.783	673	2.828	741	2.870	810	2.908	909	2.959
343	2.535	409	2.612	475	2.677	543	2.735	609	2.785	675	2.829	743	2.871	813	2.910	912	2.960
345	2.538	411	2.614	477	2.679	545	2.736	611	2.786	677	2.831	745	2.872	816	2.912	915	2.961
347	2.540	413	2.616	479	2.680	547	2.738	613	2.787	679	2.832	747	2.873	819	2.913	918	2.963
349	2.543	415	2.618	481	2.682	549	2.740	615	2.789	681	2.833	749	2.874	822	2.915	921	2.964
351	2.545	417	2.620	483	2.684	551	2.741	617	2.790	683	2.834	751	2.876	825	2.916	924	2.966
353	2.548	419	2.622	485	2.686	553	2.743	619	2.792	685	2.836	753	2.877	828	2.918	927	2.967
355	2.550	421	2.624	487	2.688	555	2.744	621	2.793	687	2.837	755	2.878	831	2.920	930	2.968
357	2.553	423	2.626	489	2.689	557	2.746	623	2.794	689	2.838	757	2.879	834	2.921	933	2.970
359	2.555	425	2.628	491	2.691	559	2.747	625	2.796	691	2.839	759	2.880	837	2.923	936	2.971
361	2.558	427	2.630	493	2.693	561	2.749	627	2.797	693	2.841	761	2.881	840	2.924	939	2.973
363	2.560	429	2.632	495	2.695	563	2.751	629	2.799	695	2.842	763	2.883	843	2.926	942	2.974
365	2.562	431	2.634	497	2.696	565	2.752	631	2.800	697	2.843	765	2.884	846	2.927	945	2.975
367	2.565	433	2.636	499	2.698	567	2.754	633	2.801	699	2.844	767	2.885	849	2.929	948	2.977
369	2.567	435	2.638	501	2.700	569	2.755	635	2.803	701	2.846	769	2.886	852	2.930	951	2.978
371	2.569	437	2.640	503	2.702	571	2.757	637	2.804	703	2.847	771	2.887	855	2.932	954	2.980
373	2.572	439	2.642	505	2.703	573	2.758	639	2.806	705	2.848	773	2.888	858	2.933	957	2.981
375	2.574	441	2.644	507	2.705	575	2.760	641	2.807	707	2.849	775	2.889	861	2.935	960	2.982
377	2.576	443	2.646	509	2.707	577	2.761	643	2.808	709	2.851	777	2.890	864	2.937	963	2.984
379	2.579	445	2.648	511	2.708	579	2.763	645	2.810	711	2.852	779	2.892	867	2.938	966	2.985
381	2.581	447	2.650	513	2.710	581	2.764	647	2.811	713	2.853	781	2.893	870	2.940	969	2.986
383	2.583	449	2.652	515	2.712	583	2.766	649	2.812	715	2.854	783	2.894	873	2.941	972	2.988
385	2.585	451	2.654	517	2.713	585	2.767	651	2.814	717	2.856	785	2.895	876	2.942	975	2.989
387	2.588	453	2.656	519	2.715	587	2.769	653	2.815	719	2.857	787	2.896	879	2.944	978	2.990
389	2.590	455	2.658	521	2.717	589	2.770	655	2.816	721	2.858	789	2.897	882	2.945	981	2.992
391	2.592	457	2.660	523	2.719	591	2.772	657	2.818	723	2.859	791	2.898	885	2.947	984	2.993
393	2.594	459	2.662	525	2.720	593	2.773	659	2.819	725	2.860	793	2.899	888	2.948	987	2.994
395	2.597	461	2.664	527	2.722	595	2.775	661	2.820	727	2.862	795	2.900	891	2.950	990	2.996
397	2.599	463	2.666	529	2.723	597	2.776	663	2.821	729	2.863	797	2.901	894	2.951	993	2.997
399	2.601	465	2.667	531	2.725	599	2.777	665	2.823	731	2.864	799	2.903	897	2.953	996	2.998
				533	2.727					733	2.865					1,000	3.000

CHAPTER VIII.

ACCURACY.

CHAPTER VIII.

ACCURACY.

INTRODUCTORY.

THERE is not in existence a perfectly accurate measurement, physical or economical, just as there is no perfectly straight line or perfect fluid. We can best illustrate the nature of economic measurements by considering that of physical. It is easy to weigh substances accurately to 1 gram: then by obtaining a good balance, we can, as our apparatus is improved, weigh accurately to a centigram, milligram, and one-tenth of a milligram; but for accuracy beyond this the balance fails us. Similarly in measuring angles, the naked eye can distinguish an object which subtends one-thirtieth of a degree; with a sextant a measurement can be taken correctly to fifteen seconds of arc; the Greenwich astronomers can make observations correct to one-hundredth part of a second, but we again come to a point beyond which precision is unattainable.

The nature of measurement.

In such cases the result is stated as correct to a milligram, or whatever it may be; in the same way we speak of an estimated sum of money correct to a pound.

A task which has considerable resemblance to some statistical estimates, is the measurement of the parallax of the sun, which determines its distance from the earth. During the eighteenth century astronomers estimated it as 10″, equivalent to 96,000,000 miles. As methods of observation and instruments were improved, observers began to agree that the whole number of seconds was 8, but gave various estimates for the first decimal figure. Since 1865 there have been very few estimates which have not given 8 as the nearest figure for this place (8.8″), while more recent observations agree in making the parallax from 8.76″ to 8.78″. We may, therefore, consider that the distance is now accurately known to within 1 in 400. Notice in this connection, first, that the earlier observations have been subject to corrections;

Physical and statistical measurements.

secondly, that better agreement has been attained as time has gone on; thirdly, that neither absolute agreement nor absolute accuracy have yet been obtained. So it is with statistical measurements; we might instance the gradual settlement of the curve representing expectation of life, the measurement of the fall in prices, and the development of wage statistics.

Degrees of possible accuracy.

Again in physical measurements, though we can sometimes reach a very high degree of accuracy, as, for instance, in the weight of a cubic foot of water which could doubtless be known correctly to one part in a million, in other cases we are glad if we can measure to one part in ten, as, for instance, in the distance of the nearest fixed star from us, which is, roughly, from 34 to 37 billion miles. So in statistics it is something if we know that the total capital of the United Kingdom was between $7\frac{1}{2}$ and 10 thousand million pounds in 1885, or if we know that the average weekly wage of workingmen in full work was from 21s. to 27s. in 1886. The weak point in such statements is that often when we have made an estimate, which we know to be inexact, we are not able to give any estimate of the limits of the error. We are not so definite as *The Modern Traveller* who

> ". . . knew the weather to a T,
> The longitude to a degree,
> The latitude exactly."

We are not able to say "our estimate is 24s. 5d., we are not certain to 1d., but it is not possible that we are as much as 1s. wrong"; whereas in physical measurements we can often give the result correct to the smallest graduation of the instrument employed.

The accuracy generally needed.

On the other hand, though we cannot obtain exactness, we can in many cases estimate to that degree of accuracy which is required for practical purpose. In common use only a certain conventional accuracy is needed. Thus, to take some miscellaneous instances, the area of an estate is given in acres, roods, and poles, but not correct to square yards; the market prices of shares do not change less than $\frac{1}{16}$; we keep the day, not the hour, of our birth; railway time-tables do not show seconds; ocean steamers are timed to start at certain hours, not minutes; height is measured correct to one-tenth of an inch; a hundred yards race is timed to one-

tenth of a second. Similarly in statistical estimates, we seldom need that our results shall be accurate within one per thousand, or even 1 per cent. One per thousand of the working week is only three minutes; 1 per cent. of the week's wage is only 3d. We do not care to know the population of London within 100, the expenditure of the Exchequer within £1,000, or the expectation of life within a day. It is often possible to attain practical accuracy within such limits.

DEFINITION OF ERROR.—For purposes of measurement we may take the following definition:—*The error in an estimate is the ratio of the difference between the estimate and the true value, to the estimate;* the error is to be reckoned positive when the true value exceeds the estimate.

Thus if the average weekly wage of agricultural labourers was in reality 14s., and we estimated it as 13s., our error would be $\frac{14-13}{13}=\frac{1}{13}$, or 7.7 per cent.; if we had estimated it as 15s., the error would be $\frac{14-15}{15}=-\frac{1}{15}$, or −6.6 per cent.

In algebraic notation, if u be the measurement of a quantity whose true value is u^1, then $\frac{u^1-u}{u}$ is the error in the estimate, which we shall call e; so that $e=\frac{u^1-u}{u}$, and $u^1=u\,(1+e)$.* $\frac{1}{e}$ is an appropriate measure of the accuracy or precision of an estimate, becoming infinite when the error is zero.

Statement of errors.

In the nature of things, when we are dealing with errors, we do not know their magnitude; the most we can know is their probable and possible extent. We might estimate, for instance, the percentage of unemployed in a certain year as 4.5, and add, from information in our possession (coming from a study of wage-bills or the reports of relief agencies), that we considered this to be within .5 of the fact; we should then write the number $4.5 \pm .5$, meaning that the error in the estimate as defined above was unlikely to be more than $\frac{.5}{4.5}=\frac{1}{9}$, or 11 per cent., and the

* This and most of the following algebraic paragraphs are from a paper on the *Relations between the Accuracy of an Average and that of its Constituent Parts*, by the present author, in the *Statistical Journal* of December 1897.

precision was 9. In such a case we can also give definite limits. The percentage unemployed must lie between 0 and 100; and if we could actually enumerate 1 per cent. of the working-class as out of work, and also 92 per cent. as in work, we should know that the number required was between 1.0 and 8.0 per cent., and the maximum error in our estimate, 4.5, was $\frac{3.5}{4.5}=\frac{7}{9}$, or 78 per cent. Even this is more precise than the original statement, "the percentage is 4.5, error unknown." By further investigation we might perhaps bring the limits of error nearer to each other, and decide that it was practically certain that the percentage required was between 4 and 4.5; then we ought to say "the number unemployed is .04 . . . of the working-class, the estimate being correct to the last figure given." This statement is of the same nature as, "The body weighs 15 lbs. 3 oz., correct to an ounce."

While, on the one hand, it is clear that we cannot often obtain close definite limits to our errors, on the other we can very often see that some of the digits in a total are almost certainly right and others almost certainly wrong. Thus when we see in the Registrar-General's Report that the population of the United Kingdom in 1895 was 39,124,496, the estimate being made from the census of 1891, and the increase calculated on the basis of the increase since 1881, we may be certain that the last two, or the last three, digits are no better than guesswork; while the first two, or the first three, are correct. Thus the statement should read: Population was 39.1 millions, or 39,124,000 ± 5,000, or whatever figures our examination of the varying rate of progress of the population led us to adopt, and this statement is actually more correct than the previous one.

Neglect of minutiæ.

It is the custom in many classes of estimates to give the figures to the uttermost farthing. This is possibly right in official publications; for the business of the office is to receive and tabulate returns, stating how and whence they came, and leaving to the economist or the statistician the task of deciding the degree of accuracy pertaining to them. But in summary descriptions and accounts, and in scientific estimates, it is not merely unnecessary to give these last figures (both because they are not accurately known, and because they generally have no importance to the argument or significance to the reader), but it is positively inaccurate.

The easiest way to avoid the inaccuracy is simply to state totals in so many thousands (*e.g.*, the earth is 8,000 miles in diameter), or if for any reason more exact measure be required (as when we are comparing the equatorial diameter with the smaller one through the poles), the scientific way is to give the number as far as it has been fairly calculated, and to indicate its precision.

Rules for Computing the Effect of Errors.

We may now give some rules connecting the errors of a complex estimate with those of the elements which form it.

I. *The error in an estimated sum is equal to the sum of the errors in the parts when each is multiplied by the ratio of the corresponding part to the sum.*

Error in sum.

For if we estimate n quantities as u_1, u_2 . . . u_n, and their sum as u, so that $u = u_1 + u_2 + \ldots u_n$, and the errors of the quantities are e_1, e_2 . . . e_n, and that of the sum is e; then the true value of the sum is $u\,(1+e)$, and the true values of the parts are $u_1\,(1+e_1)$, $u_2\,(1+e_2)$ . . . , so that—

$$u\,(1+e) = u_1\,(1+e_1) + u_2\,(1+e_2) + \ +,$$

but $$u = u_1 \qquad + u_2 \qquad + \ +;$$

hence, by subtraction, $$ue = u_1\,e_1 \qquad + u_2\,e_2 \qquad + \ +,$$

and $$e = e_1 \times \frac{u_1}{u} \qquad + e_2 \times \frac{u_2}{u} \qquad + \ +.$$

The formula is easily adapted to the case where some of the parts are subtractive.

To take an arithmetical example, if two trade unions return respectively 555 and 45 members as out of work, while the true numbers are 565 and 50, so that the errors are $\frac{2}{111}$ and $\frac{1}{9}$, then the error in the sum is by the above rule—

$$\tfrac{2}{111} \text{ of } \tfrac{550}{600} + \tfrac{1}{9} \text{ of } \tfrac{45}{600} = \tfrac{1}{40}, \text{ or } 2\tfrac{1}{2} \text{ per cent.}$$

The greater error in the returns of the smaller union has little effect on the total.

We can apply the rule to the important case where we can estimate a great part of a required total with considerable

accuracy, while we are ignorant of a smaller part. Thus we may receive returns from several unions that 33,650 are out of work, and have reason to know that the error is not more than 1 per cent., while some smaller unions do not send any returns; we make an estimate for the smaller unions, say that 1,000 of their members are unemployed, and suppose a very large error, say $\frac{2}{3}$ or 67 per cent. Then the error in the total is less than—

$$\frac{1}{100} \text{ of } \frac{33650}{34650} + \frac{2}{3} \text{ of } \frac{1000}{34650} = 2.9 \text{ per cent.},$$

an error very much nearer that of the larger returns than that of the smaller. In the preceding sentence we say "less than," because we assume that we have taken an outside limit for the smaller error.

II. *The error in the arithmetic average of several estimates is the sum of the errors of these estimates, when each is multiplied by the ratio of the corresponding estimate to that of the sum of the estimates.*

Error in average.

For if m_1, m_2, . . . m_n are n estimates of quantities whose true values are $m_1\ (1+e_1)$, $m_2\ (1+e_2)$, . . . , the estimated and true averages are respectively—

$$\frac{m_1+m_2+\ \ldots\ m_n}{n} \quad \text{and} \quad \frac{m_1(1+e_1)+m_2(1+e_2)+\ \ldots\ +m_n(1+e_n)}{n}$$

and the error in the average is—

$$\frac{\frac{m_1(1+e_1)+m_2(1+e_2)++}{n} - \frac{m_1+m_2++}{n}}{\frac{m_1+m_2++}{n}} = \frac{e_1m_1+e_2m_2++}{m_1+m_2++}$$

$$= e_1 \times \frac{m_1}{\text{S}.m} + e_2 \times \frac{m_2}{\text{S}.m} \times$$

where S denotes the sum of all the m's.

It is easily seen that no individual error can have much influence on the result, that the error in the average would be nearly of the same magnitude as one of the individual errors, if these were not very unequal, and all positive or all negative, and that if, as is generally the case, some are positive and some negative (a point we shall consider presently), the error would be considerably lessened.

III. *The error in a weighted average is the sum of* (1) *an error due to errors in the quantities, similar to the error of an unweighted average, and* (2) *an error due to errors in the weights, which becomes very small when the original quantities are nearly equal.*

Error in weighted average.

For if w_1, w_2 . . . be estimated weights applied to estimated quantities m_1, m_2 . . . , and if the true values of the weights are $w_1(1+\epsilon_1)$, $w_2(1+\epsilon_2)$ . . . , and of the quantities $m_1(1+e_1)$, $m_2(1+e_2)$ . . . , then the error is—

$$\left[\frac{S\{m(1+e).w(1+\epsilon)\}}{S\{w(1+\epsilon)\}} - \frac{Smw}{Sw}\right] \div \frac{Smw}{Sw}$$

If we simplify this expression and neglect the products of two of the errors e and ϵ (for if e and ϵ are each .1, their product is only .01), we obtain—

Error in weighted average is—

$$\left[e_1\frac{m_1w_1}{\Sigma mw} + e_2\frac{m_2w_2}{\Sigma mw} + + \right] + \left[\frac{w_1w_2(m_1-m_2)(\epsilon_1-\epsilon_2)}{\Sigma w.\Sigma mw} + \text{as many similar terms as there are pairs of quantities}\right]$$

If $m_1 - m_2$ is small, or if any other pair of the original quantities are nearly equal, the first term in the second bracket becomes very small. Very great errors are required in the weights to make any appreciable error in the average. In fact, the *errors in the quantities have so much more influence than those in the weights on the weighted average of not very unequal quantities, that errors in the weights can generally be neglected.* Many numerical examples of this principle were given in the section on weighted averages.

IV. *The error in a product is approximately the sum of the errors in its factors, due regard being paid to sign.*

Error in product.

For if f_1, f_2, . . . f_n are the estimated factors, whose true values are $f_1(1+e_1)$, $f_2(1+e_2)$, . . . , then the error of the product

$$\frac{f_1(1+e_1).f_2(1+e_2).\ldots - f_1.f_2.\ldots}{f_1.f_2.\ldots}$$

$= (1+e_1).(1+e_2)\ldots - 1 = e_1 + e_2 + + e_n$, if we neglect products of two or more e's.

The e's are equally likely, *à priori*, to be positive or negative. If two e's are of different signs, they tend to neutralise one another. The error in a product may be great if all the errors of the factors are of the same sign, even if they are small individually.

For example, if we estimate that 100 men are earning on the

average 25s. each, while in reality there are 105 men earning 26s., the error in the estimated total sum earned is, by formula, $\frac{5}{100}+\frac{1}{25}=.09$.

If, with the same estimates, the real quantities had been 105 and 24s., the error in the product would have been $\frac{5}{100}-\frac{1}{25}=.01$.

V. *The error in a ratio is approximately the difference between the errors in its two terms, due regard being had to sign.*

Error in ratio. For if u_1, u_2 be the estimated terms, whose true values are $u_1(1+e_1)$ and $u_2(1+e_2)$, then the error in the ratio is—

$$\frac{\frac{u_1(1+e_1)}{u_2(1+e_2)}-\frac{u_1}{u_2}}{\frac{u_1}{u_2}}=\frac{1+e_1}{1+e_2}-1=\frac{e_1-e_2}{1+e_2}$$

$$=(e_1-e_2)(1-e_2+e_2^2+e_2^3+-)$$

$$=e_1-e_2,$$ if we neglect terms of the second order in the e's.

If the errors in the terms are both positive or both negative, they tend to neutralise one another; if they are also nearly equal, the error in the ratio becomes very small.

We can apply Rule V. to the error in comparison of two averages of similar quantities estimated at different dates.

With the same notation as under Rules II. and III., using m, e, ϵ, for the quantities at one date, and m^1, e^1, ϵ^1, for similar quantities at another date, then the error in the ratio of the simple average of m_1^1, m_2^1 . . . to the simple average of m_1, m_2 . . . is—

$$S\left\{e^1\left(\frac{m^1}{Sm^1}\right)\right\}-S\left\{e\left(\frac{m}{Sm}\right)\right\}$$

$$=\left(e_1^1.\frac{m_1^1}{Sm^1}-e_1.\frac{m_1}{Sm}\right)+\left(e_2^1.\frac{m_2^1}{Sm^1}-e_2\frac{m_2}{Sm}\right)++$$

Now if the quantities have not changed much during the period between two observations, the fraction $\frac{m^1}{Sm^1}$ will differ little from $\frac{m}{Sm}$, and so on.

Neglecting these differences in comparison with the quantities themselves, a legitimate process when we are estimating the approximate influence of errors, we have—

$$\text{Error in the ratio of the simple averages}=S\left\{\frac{m_1^1}{Sm^1}(e_1^1-e_1)\right\}$$

If the two estimates have been made under nearly similar circum-

stances, leading to similar chances of errors, e_1^1 and e_1 are likely to be not only of the same sign, but nearly equal.

Write d_1, d_2 . . . for $(e_1^1 - e_1)$, $(e_2^1 - e_2)$ . . . , and we have—

$$\text{Error} = \text{S.}\left\{d_1 \cdot \left(\frac{m^1}{\text{S}m^1}\right)\right\}, \text{ where the } d\text{'s may be small.}$$

The corresponding analysis for the error in the ratio of two weighted averages is too complicated to be given here;* but using the principle that errors in weight are less important than errors in quantity, which applies with slight modifications, we may use the formula just given for the first approximation to the error in the ratio of two weighted averages. This formula may be put in words:—

VI. *The error in the ratio of two averages of similar series of quantities, estimated at different dates, is approximately equal to the sum of the differences between the errors in the corresponding terms of the two series, each multiplied by the ratio of the latter of these corresponding terms to the sum of all the terms at the latter date.*

Error in comparison of averages.

This rule is so important that it will be worth while to illustrate it by an example, in which a further quantity will be introduced.

If in each of two years we are able to estimate, as in our example under Rule I., one part of a total more accurately than another part, we can use the following formulæ:—

	First Year.	Second Year.
Estimated numbers or weights -	w; error ϵ;	w^1; error ϵ^1
Estimated average income, or quantity - - - - -	m_1; error e_1;	m_1^1; error e_1^1
Estimated number, less accurately known - - - - -	rw; error in r, ρ;	r^1w^1; error in r^1, ρ^1
Estimated income - - -	m_2; error e_2;	m_2^1; error e_2^1

e_1 and e_1^1 are, by hypothesis, less than e_2 and e_2^1.

Error in average for first year—

$$\frac{w(1+\epsilon).m_1(1+e_1)+r(1+\rho).w(1+\epsilon).m_2(1+e_2)}{w(1+\epsilon)+r(1+\rho).w(1+\epsilon)} - \frac{wm_1+rwm_2}{w+rw}$$

$$\frac{wm_1+rwm_2}{w+rw}$$

$$= e_1 \frac{m_1}{m_1+rm_2} + e_2 \frac{rm_2}{m_1+rm_2} + \rho \frac{r}{1+r} \cdot \frac{m_2-m_1}{m_1+rm_2}$$

if we neglect products of e and ρ.

* It will be found in the article cited on p. 201 above; a further approximation also for the error in the ratio of simple averages is there given.

Here the errors, e_2 and ρ, connected with the less accurately known part, are each multiplied by r, the ratio of the weight of that part to the weight of the better known part; while e_1, the remaining error, is by hypothesis small.

If for simplicity of argument we assume that the ratio of the unknown part to the whole (but not the error in estimating it) has remained unchanged, and also that the ratio of the estimated average incomes of the two parts has not altered, we have for the error in comparison—

$$(e_1^1 - e_1) \cdot \frac{m_1}{m_1 + rm_2} + (e_2^1 - e_2) \cdot \frac{rm_2}{m_1 + rm_2} + (\rho^1 - \rho) \frac{r}{1+r} \cdot \frac{m_2 - m_1}{m_1 + rm_2}$$

Thus in estimating the change in average wages of Scotch agricultural labourers, we have figures similar in character to the following:—

1867. MARRIED PLOUGHMEN.				1892. CORRESPONDING NUMBERS.	
Estimated number -	1,000	Average income,	£36	1,200	£49 0 0
Supposed true number	1,010	,, ,,	35	1,220	48 0 0
FARM-SERVANTS.					
Estimated number -	200	Average income—		240	
		Money -	£21		£27 5 0
		Estimated value of board -	13		14 0 0
		Total -	£34		£41 5 0
Supposed true number	220	Total income -	£37	240	£47 0 0

Here $w = 1{,}000$, $m_1 = 36$, $r = \frac{1}{5}$, $m_2 = 34$, $w^1 = 1{,}200$, $m_1^1 = 49$, $r = \frac{1}{5}$, $m_2^1 = 41\frac{11}{40}$, $\epsilon = \frac{1}{100}$, $e_1 = -\frac{1}{56}$, $\rho = -.09$, $e_2 = \frac{3}{34}$, $\epsilon^1 = \frac{1}{61}$, $e_1^1 = -\frac{1}{49}$, $\rho^1 = \frac{2}{61}$, $e_2^1 = \frac{229}{16511}$.

Here it is supposed that we have overvalued the income of the married ploughmen, and undervalued that of the farm-servants in both cases. We suppose, as is the fact, that the value of the board and other perquisites of the farm-servants cannot be estimated with precision, and that the proportionate numbers in the two classes are not accurately known.

Substituting in the above formula we find that the error in the estimated ratio of the average incomes of the two classes together in the two years is—

− .006, due to errors in estimates of income of ploughmen.
+ .008, ,, ,, ,, servants.
− .001, ,, ,, ratios of the numbers in the two classes.

Thus the last error, due to weights, is very small, and the second error, due to ignorance of the value of board, is reduced by the smallness of the number employed to a magnitude comparable with the first.

The whole error is, therefore, by formula +.001. Going to the actual figures, we find the estimated ratio of the second to the first to be 1.338 to 1, and the supposed true ratio to be 1.335 to 1; that is, the error is $\frac{.003}{1.338} = .002. \ldots$

The difference between the two methods of calculation is then 1 in the third decimal place, which is accounted for by the neglect of the less important terms.

It is to be noticed that the error in the ratio of two quantities is not the same as the error which we might be inclined to estimate, the error in the percentage increase. Thus in the case just taken, the estimated and true percentage increases are 33.8 and 33.5, and the error in the percentage increase is .01. For accuracy in such calculations, then, we require the error found by formula, according to Rule VI., to be very small.

Biassed and Unbiassed Errors.

In the consideration of all errors in averaging or comparing, it is important to distinguish two classes of errors, those which are biassed and those which are unbiassed. The difference can be made clear by illustrations. If a number of men are sent to investigate the condition of an industry in different places, with a view of proving that wages are high, conditions of work healthy, and so on, they would probably, by examining only the best conducted works, and taking the wages only of the more skilled and regular workmen, produce an average for each town which would be too high. On the other hand, if there was no brief to be held, but the investigation was impartial, the commissioners would in some towns take too high an average, in others too low, according to their idiosyncrasies and to circumstances. In the first case, the errors would be biassed, all in the same direction, all tending to increase the average, whose errors would be equal to the average error in the different towns. In the second case, the errors would be unbiassed, just as likely to be in excess or defect, and

Errors are biassed or unbiassed.

the more estimates made, the smaller would the resulting error be. The following figures would illustrate this:—

	Fact.	Biassed Estimate.	Unbiassed Estimate.
	s.	*s.*	*s.*
Average Wages in District—*a*	24	25	24
,, ,, *b*	23	25	25
,, ,, *c*	26	27	25
,, ,, *d*	27	28	28
,, ,, *e*	28	30	27
Averages - - - - -	25.6	27	25.8
Errors - - - - -	...	5·5 %	1 %

In measuring the distance of a bicycle ride on a mile-stoned road, it is found that the distances between successive milestones are not exact, but perhaps 100 to 200 yards out; but it is nearly as likely that the errors will be in excess or defect, and the greater the distance gone the smaller will be the error, as defined. The errors are unbiassed. If, on the other hand, the bicyclist trusts to his cyclometer, he will have to deal with a biassed error, for the instrument will not fit the wheel exactly, but will always register say 1,800 yards when the machine has gone a mile. This is a case where the bias can be measured and allowed for, whereas the unbiassed errors must be left to eliminate themselves. It is frequently the case that biassed errors are due to a wrongly graduated instrument; unbiassed to separate faulty measurements.

In the census returns, the fact that many women return themselves as younger than their birth certificate states, causes a biassed error in the average age of the population; the fact that people frequently return their ages at the nearest round number causes unbiassed error, and on the whole affects the average little. It is not improbable that in the Wage Census of 1886-89, there was a general tendency to obtain returns from the more liberally and better conducted establishments; this causes a biassed error in the average obtained. With these illustrations we can pass on to another principle of great importance. **Unbiassed errors are of little importance compared with biassed errors in a simple estimate; but biassed errors diminish when the ratio of two similar estimates is taken.**

Relative importance of biassed and unbiassed errors.

c

For in an average of several quantities, which have biassed errors (η_1, η_2 . . .) and unbiassed errors (e_1, e_2 . . .), it is easy to see from Rule II. that the resulting error may be written $S\left(e_1 . \frac{m_1}{S.m}\right) + S\left(\eta . \frac{m_1}{Sm}\right)$.

In the first term, the errors being unbiassed, many of them are positive, many of them negative, and they tend to neutralise one another; in fact, if E is typical of the errors e_1, e_2 . . ., then a first approximation to the error arising from them in the average is $\frac{E}{\sqrt{n}}$*.

Thus in the average of one hundred measurements, whose individual unbiassed errors are about $\frac{1}{10}$, the resulting error is $\frac{1}{10} \div \sqrt{100} = \frac{1}{100}$. There is no counterbalancing tendency, on the other hand, in the biassed errors; if each estimate was 10 per cent. in excess, then the average is also 10 per cent. in excess.

Great effect of biassed errors. When aiming at accuracy our principle always is to take care of the pounds, and let the pence take care of themselves; and it is quite futile to diminish the unbiassed errors, that is to increase the precision of our measurements, while a large biassed error runs through them all. If we do not know of the existence of biassed errors, which in reality pervade our estimates, there is no remedy; if we do know of them, we are likely to obtain more accuracy by the most erroneous corrections for them than by neglecting them; for when we make unbiassed corrections for our biassed errors, we reduce them to unbiassed errors, and then the more terms we include in our average the smaller is our resulting error. If, for instance, we find that the average weekly wage of agricultural labourers throughout the country is 13s., and by considering the circumstances of the thousand returns which we may suppose led to this average we have reason to suppose that an error of 1s. would be typical of the unbiassed errors in them, then an error of $\frac{1s.}{\sqrt{1000}}$,† that is only $\frac{2}{5}$d., may be expected to result in the average. We have here a totally illusive accuracy;

---

* See article cited p. 201, *supra*, and Part II., Sect. V., *infra*.

† More correctly the error in the average is as likely as not to be as great as this, and very unlikely to be much greater.

the part of the labourer's income which we have not included, payments at haytime and harvest, facilities for piece-work, cheap rent for cottage and land and smaller perquisites, is not capable of exact calculation. If we omit all these entirely we shall leave an error in our average of 2s. or so; but we make *individual* estimates of these additions, in all the thousand cases, though each estimate may be 2s. wrong, if there is no bias, the resulting error on the average may be expected to be $\frac{2s.}{\sqrt{1000}}$, that is only $\frac{4}{5}$d.: our whole error is now not far from 1d., instead of 2s. In estimating the accuracy of published averages, these principles should be always borne in mind, and the possibility of biassed errors always considered.

Accuracy of comparisons.

When we are dealing with the errors of a ratio the case is quite different. The error of a ratio is approximately equal to the difference between the errors in its terms; if η, η^1 and e, e^1 are the biassed and unbiassed errors in the terms, then by Rules I. and V. $(\eta^1-\eta)+(e^1-e)$ is the error in the ratio. Now the unbiassed error (e^1-e) is likely to be of nearly the same magnitude as either e or e^1;* if, as in the above example, e and e^1 are unlikely to be much greater than $\frac{2}{5}$, (e^1-e) would be unlikely to be much greater than $\frac{3}{5}$. But $(\eta^1-\eta)$, the result of the biassed errors, will, if the bias in both terms of the ratio was in the same sense (positive in both, or negative in both), be less than the original errors. If we have made the estimates of both terms on precisely similar methods, if we have asked the same questions of the same classes of persons, included and omitted the same details on both occasions, we shall have made the same errors of bias in both estimates. To return to our previous illustration, if we have made the glaring mistake of omitting everything except average weekly wages in the income of an agricultural labourer on both occasions, the only resulting error in the ratio will be that due to the *change* in these extra payments, which in short periods is likely to be small. Or, if we had taken summer wages as the

---

* If E is the probable error in e or e^1, then $E.\sqrt{2}$ is the probable error in their difference. See p. 305, *infra.*

average for the year in both cases, the error in the ratio will depend only on the *change* in the relation of summer wages to that average. Hence the error in the ratio of two estimates at different dates of a slowly changing quantity is, if the estimates are made on similar methods, often much smaller than the error in either estimate singly; for the unbiassed error is little greater, and the more important biassed error is much diminished. We need not now know of the existence of the biassed errors; they will disappear of themselves. If we are aware that there are biassed errors, and have any means of making fairly good estimates of them, it will be worth doing; but we shall make a great mistake if we correct the bias in one year and leave it uncorrected in another. For purposes of comparison it is very seldom of much use and often of great disutility to make the later estimate more accurate than the earlier. The error resulting from unbiassed errors can indeed be diminished a little,* but the error resulting from the more important biassed errors will only be increased. All Government officials and others who compile annual returns are in a dilemma: to make their annual statements accurate in themselves, they should always be straining after improvements, they should always be watching for changes in the quantities measured and adapting their methods and tabulations to these changes; but to make their annual returns comparable with each other, they should be absolutely conservative, and cling to any mistakes they or their predecessors have made in the past with all the strength red tape can give them, being careful, however, not to add to the mistakes or make new omissions. The dilemma can in some cases be avoided; for when an improved method is introduced, the tabulation can sometimes be given for a few years both on the old and on the new plans; then when the difference introduced by the change is known, the earlier figures can be brought to the greater precision of the later. Thus the Board of Trade has recently included in the tabulation of exports ships which, leaving our shores with merchandise, are them-

Need for uniformity in structure of serial returns.

* For if E and E_1 be typical of the unbiassed errors at the two dates, then $\sqrt{E_1^2+E^2}$ is typical of the error in the ratio, which diminishes with either E or E_1. See p. 305, *infra*.

selves sold to a foreign owner; and we have the following tabulation:*—

	1899.	1898.
Exports of Home Products (exclusive of ships sold to foreigners) - - -	£255,465,000	£233,359,000
Re-exports of Home and Colonial Merchandise -	65,020,000	60,655,000
Total - -	£320,485,000	£294,014,000
Value of New Ships exported	9,195,000	Not stated.
New total -	£329,680,000	

Ignorance of slight alterations in the collection and tabulation of material has been the cause of many statistical mistakes.

To sum up the chief results of this chapter: there are two processes which tend to accuracy—*averaging*, which diminishes unbiassed errors; and *comparison*, which diminishes biassed error. *The errors in weights are seldom so important as the other errors which are present in estimates.* Errors in a result cannot, of course, be calculated, but can be expressed in terms of errors in the items, from which it comes; we cannot attain certainty, but we can indicate processes which diminish errors, and with the help of mathematics measure the extent of diminution. Initial errors are diminished most, when we calculate the *ratios of weighted averages of similar and similarly estimated quantities.* Index-numbers, which we discuss in the next chapter, are examples of this class.

Results.

The accuracy resulting from the process of sampling requires more mathematical treatment, and is dealt with in Part II., Section V.

* Quoted from the *Economist*, 17th February 1900, in the *Statistical Journal* of March 1900.

CHAPTER IX.

INDEX-NUMBERS.

CHAPTER IX.

INDEX-NUMBERS.

THE discussion of index-numbers supplies so good an illustration of the principles laid down in the last chapter, and index-numbers are so important in themselves, that, though it is our intention to avoid special questions, it will be worth while to devote a short chapter to them.

Index-numbers are used to measure the change in some quantity which we cannot observe directly, which we know to have a definite influence on many other quantities which we can so observe, tending to increase all, or diminish all, while this influence is concealed by the action of many causes affecting the separate quantities in various ways. Thus, to take three of the quantities to which index-numbers are applied, the change in the relation of the precious metals to the work to be done by them affects prices of all commodities, but very many other causes are at work affecting the prices of separate groups of commodities; there are general causes tending to raise the wage of a week's work of average skill, but this general increase is concealed by numberless minor causes affecting different grades of labour in different degrees; the change in the consumption of goods by the working or other classes is a sufficiently definite quantity, but it can only be measured indirectly by observing the varying changes in the consumption of individual articles.

Function of index-numbers.

The use of index-numbers is not, however, confined to these instances, but is nearly co-extensive with the field of statistics; for we have limited the term statistics to the measurement of complex groups and their changes; the object of statistics is to measure the action of the general laws which govern a heterogeneous group, and the changes produced by general forces can be measured, as a rule, only by their effect in individual cases; thus the method of index-numbers is at once applicable to the

disentanglement of that which is common to the whole group from those variations which are special to individual items.

The general method of forming an index-number, *e.g.*, of the fall of prices, is as follows:—We select commodities, whose prices we can estimate accurately, and tabulate their prices for a series of years. Choosing the prices in one year or the average for a sequence of years as a base, we express each series of prices as percentages, year by year, of their height in the chosen base-year, or their average in the chosen period. Then to find the index-number for any year in particular we take the average of the percentages in that year.

Method of formation.

The problem, of which index-numbers should give the numerical solution, may be compared to that presented to astronomers who estimate the motions of the sun by observing those of the stars. As the sun and earth move towards some distant point, say in the constellation Hercules, the stars have an apparent motion, due to the unperceived motion of the observer; those in the region of space towards which he is travelling appear to be spreading out, as the distances separating them gradually subtend wider angles, while those in the region from which he is moving appear to close together, and those in directions perpendicular to the line of movement appear to move backward. Meanwhile all these stars have their proper motions, as rapid as that of the sun, but in as many different directions as there are stars. On the whole there is a trend in the directions determined by the sun's motion, but in individual cases this trend is entirely lost. So when a change in the currency has a general influence on prices, this influence is concealed by the movements due to causes affecting only some of the commodities. In both cases it is possible to find the general trend, if sufficient accurate observations are available. In both cases the problem is complicated by the possibility of links connecting the movements of groups of the stars or of the prices.

Astronomical analogy.

It has sometimes been supposed that we can estimate the effects of general causes directly; that we can, for instance, obtain an objective measurement of the change in the purchasing power of gold, by evaluating it at two dates in terms of all commodities purchased, weighted by the amount spent on each; but it is better to neglect this method at once both as impracticable and as not answering the purpose of index-

Index-numbers and samples.

numbers, for the effects of minor causes affecting separate commodities would not then be necessarily separated from the main cause.

Suppose that the changes in a group of quantities are determined by one general force which acts on all in the same sense, that is, tends to increase all or decrease all, and by several other forces each of which acts on one or more of the quantities, and some of which tend to increase, others to decrease the quantities they affect; then of the special forces, some will tend to increase, others to diminish the average, while the general force will have a cumulative effect entirely towards increasing, or entirely towards diminishing it. If the separate effects of the special forces are small compared with their number, they will tend to neutralise one another in their influence on the average; and the change in the average will show the influence of the general cause only.* In the language of the last chapter, the special forces produce unbiassed changes, which are negligible in their effect on an average, in comparison with the biassed changes produced by the general force. To obtain this elimination it is necessary to take random samples, so that the laws of probability may have free action; and the two questions to be discussed are the choice of samples, and the choice of weights to be applied to them.

Unimportance of weights.

As we have already seen, the effect produced by varying the system of weights applied to so few as 30 or 40 numbers is very slight, and the error resulting from errors in weighting is in many cases much smaller than the error resulting from faulty measurements of the quantities weighted. We shall presently show † that the precision of an average increases with the number of like quantities averaged. From these principles it is clear that it is more important to increase the number of our samples than to attempt accurate calculations of the proper weights to give them.

The choice of samples is in practice very much limited, for in calculations extending over long periods we are dependent on the accidental preservation of records; and when we have taken

* This abbreviated statement should be criticised in the light of Part II., Sect. V., *infra*. See also Report of Committee on *Variations in the Monetary Standard*, British Association, 1888-90.

† See p. 305, *infra*.

into our reckoning all the measurements which can be accurately made, the number of samples barely comes up to the minimum necessary for the normal action of the laws of probability.

There are many index-numbers of wholesale prices extant, some of which we may pass in review. The Board of Trade publish the recorded quantity and value of goods imported and exported, and the average prices of these goods can be calculated. Those commodities are selected which occur in the returns for the whole period chosen. A particular year is chosen as base; then the goods are valued in all other years separately at their prices in the base year; the total of these values in any year is the sum which the goods would have been worth if their prices had remained unchanged; the ratio of this value to that actually recorded is the ratio of their average price in the base year to their average price in the other year selected (if the term average is used broadly), and if the first term of this ratio is equated to 100, the second term is the index-number required for the year selected, expressed as a percentage of the number for the base year. It is at once evident that we are here dealing with weighted averages.

The Board of Trade index.

Let $p_1, p_2, p_3 \ldots$ be the prices in the base year of units of the goods selected, and $r_1p_1, r_2p_2, r_3p_3 \ldots$ the prices in the year for which we require an index-number: then $r_1, r_2, r_3 \ldots$ measure the changes of prices for the separate commodities, and these r's are the samples from which we are to deduce the general change of price. The weights used in the process described may be found thus: let $b_1, b_2, b_3 \ldots$ be the numbers of units of goods in the selected year; then the total value in the selected year at the prices of that year is $(b_1r_1p_1+b_2r_2p_2+\ldots)$, and at the prices of the base year is $(b_1p_1+b_2p_2+\ldots)$; the ratio is $\Sigma brp : \Sigma bp$, and the index-number for the selected year is

$$100 \times \frac{\Sigma brp}{\Sigma bp} = 100 \times \Sigma\left(r\frac{bp}{\Sigma bp}\right).$$

Systems of weights.

Here the weights applied to the r's are the values which the corresponding goods in the selected year would have borne at the prices of the base year. It is clear that the selection of the standard year affects the weights, for any particular commodity can be given special weight by choosing as base a year in which its price is high, and much trouble has been spent in searching

for a "normal" year; but though the weights of separate commodities are affected, it does not follow that the average will be altered, and we should expect from the principle laid down above that the change would be very slight. In fact we have the following figures:—

INDEX NUMBERS OF 1886 AND 1883 COMPARED.*

Weights.	Imports. Values at 1873 Prices.	Imports. Values at 1883 Prices.	Imports. Values at 1861 Prices.	Imports. Values at 1881 Prices.	Exports. Values at 1873 Prices.	Exports. Values at 1883 Prices.	Exports. Values at 1861 Prices.	Exports. Values at 1881 Prices.
1883	100	100	100	100	100	100	100	100
1886	81.7	82.1	82.9	82.3	88½	88	87	89

It is possible to produce figures which show a variation caused by a change of base year, but it is done by choosing samples which lend themselves to the special argument.

Since so great an alteration in choice of weights makes so little difference, it is worth while to see if we need even keep the weight due to the quantities imported (the b's in the above formulæ). The following table may be quoted† to show that these weights even have practically no influence:—

Index-Numbers for 1895, *when that of* 1881 *is* 100, *obtained by Various Systems of Weighting.*

	Ratios of Prices ($r_1, r_2 \ldots$) Weighted by Values of 1895 Quantities at 1881 Prices.	Weighted by Declared Values in 1881.	Arithmetic Mean.	Median.	Geometric Mean.	Reciprocal of A.M. of $\frac{1}{r_1}, \frac{1}{r_2}$, &c.	*Economist's* Figures.
Imports	67½	69	73½	72½	72½	69	71
Exports	83	87	82	81	78½	75	

* From the *Economic Journal* and the *Statistical Journal*, both June 1897.

† From the *Economic Journal* (with a correction in the statement of weights).

In the first column of figures the goods of 1895 ($b_1, b_2, b_3 \ldots$ units at prices $p_1^1, p_2^1, p_3^1 \ldots$) are valued at the prices of similar goods in 1881 ($p_1, p_2, p_3 \ldots$). The ratio of their new to their old value $\left[\frac{\Sigma bp^1}{\Sigma bp} = \Sigma\left(r.\frac{bp}{\Sigma bp}\right)\right]^*$ is the ratio of the new index-number to 100. In the next column the index-number is obtained by valuing the quantities of 1881 at the prices of 1895; then the ratio of the new value to the old $\left(\frac{\Sigma ap^1}{\Sigma ap}$ (where $a_1, a_2 \ldots$ are quantities in 1881) $= \Sigma r.\frac{ap}{\Sigma ap}\right)$ is the ratio of the index-number of 1895 to 100 for 1881. In the next three columns the arithmetic mean, the median, and the geometric mean of the r's are given. In the last column but one the arithmetic mean of $\frac{1}{r_1}, \frac{1}{r_2} \ldots$, that is of the ratios of the prices of 1881 to 1895, is calculated, and the ratio of this mean to 100 equals the ratio of 100 to a new index-number, which corresponds to the former arithmetic mean with the years 1881 and 1895 interchanged. The figure in the last column is calculated from material given in the *Economist*;† every year the imports and exports are valued at their prices in the previous year, and thus an annual ratio is given similar to that in the first column of figures in the table just given; the number 100, taken for 1881, is multiplied by this annual ratio year by year till 1895, and the number 71 is the result. [Algebraically this index-number is—

$$100 \times \Sigma\left(r.\frac{bp}{\Sigma bp}\right) \times \Sigma\left(r^1.\frac{b^1p^1}{\Sigma b^1p^1}\right) \times \times .]$$

A more complete analysis of these figures, and an investigation as to the causes of the divergence between the export indices 87 and 75, would show which of the methods should be adopted. Here we will be content with noticing that the unweighted average, 82, is very near the first weighted average, 83.

Further methods of dealing with such weights are given on p. 226, under Retail Index-Numbers.

---

* Where $r_1 = \frac{p_1^1}{p_1}$, $r_2 = \frac{p_2^1}{p_2}$, &c.

† See, for instance, the quotation from the *Economist* in the *Statistical Journal*, 1900, p. 129.

The advantage of index-numbers on the Board of Trade basis is that they measure approximately an objective quantity, and a result is obtained which can be stated in terms which appeal to the ordinary man who is not a statistician: such as, "The imports of 1895 would have cost half as much again if their prices had been those of 1881;" but, as pointed out above, it does not follow that this index is the best measure of the less-definable quantity, "Fall in the price of imports," where we imagine a general cause affecting this class of commodities whose action is modified by other partial causes.

Objective measure.

A special advantage of the geometric mean* is that the results it gives are independent of the year chosen as base; for if $p_1, p_2, \ldots p_n$ and $p_1^1, p_2^1, \ldots p_n^1$ are the prices in two years, $\sqrt[n]{p_1 p_2 \ldots p_n} : \sqrt[n]{p_1^1 p_2^1 \ldots p_n^1} :: 100 : I_1$, the required index-number; hence—

Geometric mean.

$$I_1 = 100 . \sqrt[n]{\frac{p_1^1}{p_1} . \frac{p_2^1}{p_2} \ldots \frac{p_n^1}{p_n}}$$

$$I_2 = 100 . \sqrt[n]{\frac{p_1^{11}}{p_1} . \frac{p_2^{11}}{p_2} \ldots \frac{p_n^{11}}{p_3}} = 100 . \sqrt[n]{\frac{p_1^{11}}{p_1^1} . \frac{p_1^1}{p_1} \times \ldots}$$

$$= 100 \times I_1 \times \sqrt[n]{\frac{p_1^{11}}{p_1^1} . \frac{p_2^{11}}{p_2^1} \ldots}$$

which would be the value obtained for the third year if only the second and third were considered. Considering the extra labour involved in calculating this mean, and the small advantage obtained by any alteration in the weighting, its use is not to be generally recommended.†

Mr Sauerbeck and the *Economist* both avoid in part the difficulty of weighting the separate ratios by their relative importance in consumption, by selecting from those commodities whose prices are most accurately determined more instances of such widely consumed articles as wheat than of less important commodities such as linseed. Mr Sauerbeck has, in his annual articles in the *Journal of the Royal Statistical Society*, verified the correspondence of the unweighted average of his 56 ratios with the average of the same weighted on various principles.

Other index-numbers.

---

* Pointed out by Professor Edgeworth.

† On this point, and on others in this chapter, see article Index-Numbers, in Palgrave's *Dictionary of Political Economy*.

While the choice of the special weights to be employed is, when the number of ratios taken is at all considerable, quite unimportant, the choice of the quantities dealt with has great effect on the result. Thus import figures, relating to raw materials and the produce of other countries, do not lead to the same index-numbers as export figures dealing with the price of our own produce, though the tables just given show that they are little affected by weights; and neither of these agree closely with Mr Sauerbeck's or the *Economist's* numbers, and these again are not in complete agreement. The samples on which these four sets of numbers are based are from different groups of commodities, and the numbers show that the same forces do not affect these groups in the same degree. When we have so multiplied our samples, that we can subdivide them without affecting the index-numbers deduced, we may expect our results to represent the required measurement.*

Importance of right choice of samples.

If we compare the *Economist* index-numbers with Sauerbeck's during the period 1860-70, we see that the former show a very much greater increase during the cotton famine than the latter. An index-number which can be greatly disturbed by fluctuations, however violent, in only one group of commodities, is clearly wanting in some of the chief qualities of a general measure of price levels. A very simple means of avoiding this difficulty, and indeed all the intricacies of weighing, is to take the median of all the price ratios of a particular year as the index-number of that year. It is perhaps impossible to show theoretically that any other average satisfies the required conditions better than the median, and there can be no doubt that it is practically the easiest to calculate.

Great advantage of the median.

If, on the other hand, paucity of data makes the inclusion of weights necessary, and the popular desire for concrete measurements makes a fine show of weighting expedient, we perhaps cannot do better than to adopt the standard proposed by the Committee of the British Association, already mentioned, for the construction of an index-number,

Proposed standard.

* Mr Sauerbeck's numbers are to be found in annual articles by him in the *Statistical Journal;* and a diagram showing them from 1820 is published by Effingham Wilson (1s.).

which might be the basis of business transactions involving future payments. This standard is as follows:—

Basis of Index-Number recommended by the Committee appointed by the Economic Section of the British Association, 1888.

Articles.	Estimated Expenditure per Annum on each. 000,000's omitted.	Hence Weights assigned.	Prices to be taken from
Wheat	£60	5	Gazette average, English wheat.
Barley	30	5	〃 〃 barley.
Oats	50	5	〃 〃 oats.
Potatoes, rice, &c.	50	5	Av. import price, 〃 potatoes.
Meat	100	10	Market quotations, live meat, Smithfield.
Fish	20	2½	Board of Trade Returns; average per cwt. landed.
Cheese, butter, milk	60	7½	Cheese and butter, average import price.
Sugar	30	2½	Av. import price, refined sugar.
Tea	20	2½	〃 〃 〃 tea.
Beer	100	9	〃 export 〃 beer.
Spirits	40	2½	〃 import 〃 spirits.
Wine	10	1	〃 〃 〃 wine.
Tobacco	10	2½	〃 〃 〃 tobacco.
Cotton	20	2½	〃 〃 〃 cotton.
Wool	30	2½	〃 〃 〃 wool.
Silk	20	2½	〃 〃 〃 raw silk.
Leather	10	2½	〃 〃 〃 hides.
Coal	100	10	〃 export 〃 coal.
Iron	50	5	Market price, Scotch pig-iron.
Copper	25	2½	Av. import price, copper ore.
Lead, zinc, tin	25	2½	〃 〃 lead ore.
Timber	30	3	Average import price.
Petroleum	5	1	〃 〃 〃
Indigo	5	1	〃 〃 〃
Flax and linseed	10	3	〃 〃 〃
Palm oil	5	1	〃 〃 〃
Caoutchoux	5	1	〃 〃 〃

Since we can only obtain rough correspondence in dealing with wholesale prices, we cannot expect to be able to measure retail prices with any great precision. For we saw in the preceding chapter that the error in an average bears a definite relation to the errors in the items which compose it; if the errors in the items are on the whole doubled, it is likely that the errors in the average and in the ratio of two averages will also be doubled, and we shall need four times * as many samples to restore the precision. Unfortunately the

Retail price index.

* See p. 305, *infra*.

material for computing a retail index-number is even more incomplete than that for wholesale prices, and owing to the smaller number of articles that can be included, and the preponderance of such items as bread and rent, the question of weighting becomes of more importance.

Special difficulties.

When we wish to construct an index-number to show the purchasing power of money to special classes, we must take into account some considerations which can be ignored when dealing with wholesale price numbers. Different classes of persons at the same time, and the same classes at different times, spend their income in varying proportions on different objects. If we could collect enough sufficiently accurate samples, this fact would not matter so much; but it would still be of some importance owing to the tendency to make increased purchases of cheapening commodities. As it is, it would be necessary to construct separate index-numbers for each class and each district. The difficulty of insufficient and inaccurate data cannot at present be overcome; but as it is possible that we may in the future get definite records of retail prices sufficiently numerous to make up for their want of precision, we may glance at the other details of the problem. To form an index-number for a particular class of people, we need records of the method of expenditure of their income at all the dates in question, of sufficient numbers to obtain the slight precision which weighting needs. Then if we had fairly good records of retail prices several methods of weighting are open to us,* all of which are likely to give nearly the same result. The necessity of weighting and the methods are best shown by a numerical illustration.

Methods of weighting.

Suppose the following records of expenditure:—

First Year.	s.	d.	Second Year.	s.	d.
6 quarterns bread at 6d.	3	0	7 quarterns at 5d.	2	11
4 lbs. meat at 7d.	2	4	5 lbs. at 8d.	3	4
½ lb. tea at 3s.	1	6	1½ lbs. at 1s. 4d.	2	0
	6	10		8	3

The second year's budget at the first year's prices would cost 10s. 11d.; index-number of retail prices on this basis—

$$100 \times \frac{8s.\ 3d.}{10s.\ 11d.} = 75.6\ (a).$$

* See article on Wages, Nominal and Real, in Palgrave's *Dictionary of Political Economy*, pp. 640-641.

The first year's budget at the second year's prices would cost 5s. 10d.; index-number on this basis $100 \times \frac{5s.\ 10d.}{6s.\ 10d.} = 85.4$ (*b*).

The ratio of the costs of $6\frac{1}{2}$ quarterns, $4\frac{1}{2}$ lbs. meat, 1 lb. tea, the averages of the quantities in the two years, at the two sets of prices, is 8s. $10\frac{1}{2}$d. to 7s. $0\frac{1}{2}$d. = 100 : 73.2 (*c*).

If we disregard the records of changing expenditure, we find that the unweighted average of the three ratios of prices is 100 : 80.8 (*d*).

If we suppose the second sum (8s. 3d.) to be spent on bread, meat, and tea in the same proportional parts as the first sum (6s. 10d.), we have—

$\frac{3s.}{6s.\ 10d.}$ of 8s. 3d., *i.e.*, $\frac{1782}{41}$d. would have bought $\frac{1782}{41 \times 5}$ quarterns at the second price, which would have cost $\frac{1782}{41 \times 5} \times 6$d. at first price.

Working out the other items in a similar way, we find that the second sum distributed in the same proportions as the first would have bought goods which would have cost 10s. $10\frac{1}{2}$d. at the first prices; and the resulting index-number is—

$$\frac{6s.\ 10d.}{10s.\ 10\frac{1}{2}d.} \text{ of } 100 = 62.8 \ (e).$$

This reduction is due to the large expenditure on bread on this hypothesis, which can easily be shown to be an unreasonable one if we suppose the price of bread to be reduced to nothing, while the other prices rise; then on this hypothesis the fall is infinite.

Of the above numbers, (*c*), (*d*), and (*e*) do not seem to rest on sound hypotheses; (*a*) clearly overstates and (*b*) clearly understates the fall; and therefore some number between (*a*) and (*b*) is the number required. If (*a*) and (*b*) lie close together there is no further difficulty; if they differ by much they may be regarded as inferior and superior limits of the index-number, which may be estimated as their arithmetic mean (80.5) as a first approximation.

Further difficulties.

While it is useful to have a definite means of calculating these numbers to bring extravagant statements to a numerical test, there are two further considerations which hinder the complete solution of the problem. In all budgets rent is an important item, and there seems no prospect of obtaining any good estimate of the relation between increasing rent and improving accommodation, allowing for the

benefits of public expenditure paid by rates included in rent. Again, if we consider, not how money is spent, but how it might be spent, we should have to introduce a more general factor; for the margin which remains when necessities are satisfied has a rapidly growing purchasing power, as the products of machinery increase in variety and diminish in price; perhaps the calculated fall in wholesale prices forms a fair measure of this growth.

Leaving this somewhat unfruitful topic, let us return for a moment to the measurement of a quantity more typical of index-numbers.* If we have to measure the action of a cause, which affects quantities which have no common measure, we are still able to apply index-numbers.

Index-numbers of consumption.

A general increase has taken place in the consumption of imported goods, and if we can measure this increase independently of any change in price, we can use it as an argument to support the alleged increase in real wages. The only common measure of bread, currants, cheese, meat, &c., of practical value is their price, their weight being useless for the purpose. If the quantities consumed year by year of a number of such commodities are written down, expressed as percentages of the consumption in any years (not necessarily the same), we have series of numbers which only need weighting to form the index-number required. We can in this case verify, that any logical choice of weights, based on their value or their assumed importance, or even a random system of weights, gives much the same index-number as the simple arithmetic averages; in fact, we have a sufficiently good group of samples to render us nearly independent of weights. When this is the case we can say with safety that the number required lies in the neighbourhood of the group given by the various systems of weights, and choose what appears the most logical system for the estimate we adopt. In the paper referred to, five different systems applied to only fourteen commodities give results for the increase of consumption all between 13.8 and 20.1 per cent. in the period 1873-96.

The application of index-numbers to wage statistics does not involve any fresh principles. It is not permissible to ignore weights in this case; for an unweighted average would not allow for the general tendency to increase numbers where wages are rising. There is great liability

Wage index-numbers.

* The following illustration is based on Mr G. H. Wood's paper on *Some Statistics of Working Class Progress*, cited above.

to "biassed" errors in separate averages; for wages for overtime, specially high piece-wages, wages of large uncombined classes of low-skilled or badly paid workpeople, may often be omitted in wage records. These biassed errors, however, tend to disappear in comparison; and it may prove possible to construct a wage index-number of very fair precision.

Note.—Pages 226-7 have been misunderstood. The method has been confused with the measurement of the cost of a standard budget (representing either minimum subsistence, or efficiency subsistence) with the items the same at both dates. The cost of such a budget is very important in relation to working-class progress, but in the text we are concerned with the change of the purchasing power of money, expended on retail commodities, which makes a different problem. Of course the three commodities used in the test form only an abbreviation for a dozen or a score of items.

The methods (*c*) and (*d*) are ignored because they do not allow for the correlation between distribution of expenditure and price; and (*e*) seems contrary to experience, and is liable to great disturbance if a single article becomes very cheap or very dear. These methods have their use in other measurements by index-numbers.

LIBRARY OF THE BOMBAY BRANCH OF THE ROYAL ASIATIC SOCIETY

CHAPTER X.

INTERPOLATION.

CHAPTER X.

INTERPOLATION.

1. GENERAL.

IT is very often the case in practical statistics that we are not able to make serial estimates as frequent or descriptions of groups as detailed, as is necessary for their use in further investigations. Thus the population is only counted once in ten years; but we need to bring monthly and annual accounts—births, deaths, trade returns, &c.—into close relation to the existing number of people, and estimates for the budget and the yield of taxes must be based on the assumed number of taxpayers for the current year; it is therefore necessary to *interpolate* estimates for the number of the people in intercensal years. Again, interpolation is needed for the statement of the distribution of the population according to age, a tabulation which is necessary for actuarial work and for sociological purposes. The ages returned on the householder's schedule are nominally correct to the year, but in practice they are known to be inaccurate, tending to group themselves in the neighbourhood of round numbers; but the returns for such age periods at 35-45 years are more correct, since the persons who return themselves as 40 years old are probably within 5 years of that age. The original returns are so erroneous that they are not published at all, but the numbers are only given in the ten-yearly periods; from the numbers so given, it is necessary to estimate the numbers for the individual years. Again, the compilers of the wage census of 1886-91 enumerate the numbers earning wages "of 15s. and under 20s.," "of 20s. and under 25s.," and so on, but not the numbers in shilling limits. In problems relating to wages we often need more detail; and when we are comparing these wages with a similar group in France, we must devise a scheme by which grades of 2 francs can be compared with grades of 5s., by a suitable system of interpolation. Such a necessity is very

Necessity of interpolation.

common when we wish to compare groups, which are similar but tabulated on diverse systems. Thus, two countries conduct their census at different dates. In one country the age groups are of fifteen years, in another of ten; in one, "young persons" are those under 21; in another, those under 18. Occasional estimates seldom correspond in date; wage statistics are found for 1840, 1850, and 1892 in France, and for 1866, 1885, 1886, and 1891 in England. Similar differences are found when we are comparing county with county; and a discussion of the method of determining averages in such a case will illustrate some of the elementary problems of interpolation.

Elementary example.

Suppose that the figures printed in Roman type in the following table are accurate returns of the weekly wages in three districts, and that we wish to find the average change in the three together.

Years.	1860.	1862.	1864.	1866.	1870.	1871.	1875.	1878.	1880.	1881.
	s. d.	*s. d.*	*s. d.*	*s. d.*	*s. d.*	*s. d.*	*s. d.*	*s. d.*	*s. d.*	*s. d.*
District A	12 6	15 0	*15 0*	15 0	*15 0*	14 6	18 0	18 0	17 6	17 0
" B	18 0	*19 0*	19 0	*20 0*	20 0	*19 6*	*21 0*	21 0	*20 6*	20 0
" C	10 0	*11 0*	11 0	12 0	*12 0*	*12 0*	15 0	*15 0*	15 0	*14 6*
Average	*13 6*	*15 0*	*15 0*	*15 8*	*15 8*	*15 4*	*18 0*	*18 0*	*17 8*	*17 2*

It is clear that there is something to be learnt about the general course of wages from the data, but the lessons are not obvious. The following figures, printed in the table in italics, are those which naturally suggest themselves. There is no sign in A of any change between 1862 and 1866, so we write *15s.* for 1864. Judging from B, the figure for 1870 is not likely to have been lower than that for 1864, so we write *15s.* for A in 1870. A is now complete; we notice that in A the first rise was complete by 1862, and assuming the same in B, we obtain *19s.* for 1862. In C there is a rise between 1864 and 1866, while in A there is no change from 1866 to 1870; B will correspond if we write *20s.* in 1866. If we write for B, *19s. 6d.* in 1871, *21s.* in 1875, and *20s. 6d.* in 1880, we shall have close correspondence with A from 1866 to 1881. Similar reasons lead to the numbers interpolated for C. The unweighted average can then be cal-

culated year by year, which could not be done directly from the data. This average reflects all the changes in the original figures and gives no special predominance to any. It may be regarded as the most probable series that can be based on the given information.

Assumptions made.

We will now notice the assumptions tacitly made in proceeding by this method. First, it has been assumed that there are no sudden jumps, that such a figure as 20s. for A 1864 is inadmissible; this is only justifiable if we are acquainted with the general causes which influence the rate of wages, and know that there was no violent disturbance in the intermediate dates. We could not make this assumption as to wages in the cotton trade in the time of the American Civil Wars, nor can we make it over a long series of years. Secondly, it has been assumed that in the absence of evidence to the contrary the rise or fall has been uniform. Thus, in B 1878-81, the wage in 1880 is assumed to be intermediate between 1878 and 1881; if there had been no indication from A that it was half-way between in point of wages, it might have been said that in point of time it was two-thirds of the way, and 20s. 8d. should be interpolated for 1879 and 20s. 4d. for 1880, if it was worth while to depart from round numbers. Thirdly, it has been assumed that the course of wages in the three districts was similar. Thus in A there is a rise from 1860-62, but there is no further improvement at any rate before 1866; it is consequently assumed that the rise registered in B and C before 1864 actually took place before 1862. Again, when considering the period 1870-75, we notice that in A there is a fall till 1871, and a sharp rise to 1875, and no change to 1878; in B, therefore, it is assumed that the wage of 1875 is equal to that of 1878, and the fall in 1878 may be allowed because it increases the sharpness of the rise in 1871-75. In C it is doubtful whether the 12s. in 1871 should not rather be 11s. 6d. The reasons against are that a gain on a low wage is often not so easily lost as a gain on a high one; 6d. is a larger drop proportionately on 12s. than on 15s.; that the rise of 3s. 6d. which would then be shown 1871-75 is a larger proportionate rise than in either A or B; and that the existence of the fall in 1870-71 depends only on the evidence of a fall between 1866-71. When the figures are few in number, it is necessary to examine them in this way to pick out the most probable; and it is often

fairly easy to fill in the figures which satisfy all the existing evidence fairly closely.

The question at once arises, What certainty have we that these quantities, by hypothesis unknown, are in reality anywhere near the figures which on the face are most probable?

In some cases of interpolation, dealt with presently, the answer can be given as a statement of mathematical probability, such as: it is 2 to 1 against a divergence of 6d. from the assigned figure, 30 to 1 against one of 1s., 1,000 to 1 against one of 2s. 6d., and so on; but in the figures most often cropping up in investigations it is not possible to assign such a precise probability. There is one rough but useful way of testing the accuracy of such interpolation as in the case before us which can be explained by an example. Test how far we can throw out our calculated average for 1870, without violently infringing the common-sense of the question. Make A and C as large as possible in these dates; we may perhaps suppose a rise of 1s. above 1866, seeing that there is one in B between 1864 and 1870. We can hardly suppose either that 1870 is as high as 1875-78, or that there is a great drop of as much as 2s. in the single year, if we are acquainted with the causes that determine the wages at those dates. Let the highest wage we can assign to A and B be 16s. 6d. and 13s. 6d. respectively. Our average is then 16s. 8d. instead of 15s. 8d. Similarly, we might perhaps think that 14s. and 11s. were the lowest possible in A and C in 1870; then the average would be 15s. Assuming that we know enough about the general trend of events at these dates to assign limits in this way, we can say it appears improbable that the average wage in 1870 was less than 15s. or more than 16s. 8d., and that the evidence points to 15s. 8d.

Tests.

The accuracy of our interpolation then depends—(1) On knowledge of the possible fluctuations of the figures, to be obtained by a general inspection of the fluctuations at dates for which they are given; (2) on knowledge of the course of the events with which the figures are connected.

Numerical example.

A second example of a similar kind* may be given to illustrate the numerical calculation.

* Taken from *Agricultural Wages in England*, in the *Statistical Journal*, December 1898, by the present author.

Northern Counties.	Weekly Agricultural Wages in 1867-69. s. d.	1869-70. s. d.
Cheshire	*13 1*	13 6
Lancashire	15 0	15 0
West Riding of Yorkshire	14 6	16 5
East ,, ,,	14 6	*14 11*
North ,, ,,	14 6	15 4
Durham	16 6	16 0
Northumberland	16 6	16 7
Cumberland	*14 4*	14 9
Westmoreland	*15 7*	16 1

Roman figures given. Italic figures interpolated.

The averages of the wages in the five districts for which data exist in both periods are 15s. 4.8d. in 1867-69 and 15s. 10.4d. in 1869-70, that is in the ratio 33 : 34. If we assume that the wages in the other counties have been influenced by similar causes and increased in the same ratio, we obtain the figures interpolated in the table. The unweighted averages for the northern counties are now 14s. 11d. and 15s. 5d. in the two periods, instead of 15s. 3d. and 15s. 5d., the averages of the given numbers. For general comparison all over England between these two years we should have been obliged to neglect the missing counties in both years, which would have unfairly lowered the general average, since these counties have in recent times had wages above the English average though below that of the northern district. At the same time we should have unfairly raised the apparent average of the northern district. We should also have lost the probable figures for the special counties at the earlier date which are on a fairly safe basis; for the wages in these counties of the Northern District remain in nearly the same order through the last fifty years. At the same time it is easily seen that these wages are not so accurately known as those not interpolated, and it is well to notice in arguments based on such figures, to what extent the interpolated figures are involved.

A process very similar to that just employed is used in giving marks at school to students who are absent from a lesson; attention is paid both to the particular student's general place in the class order, and to the average value of the marks obtained by the rest of the class in the lesson missed.

Though the method be fairly complete it is very important

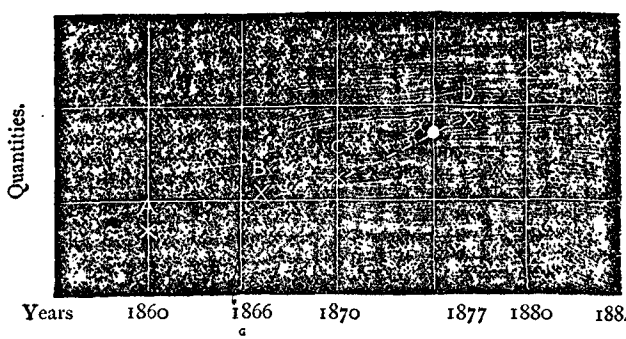
to notice that interpolated figures rest on quite a different class of evidence to those which are the result of direct evidence. In some cases they may represent quantities which have no existence (as in the case of school marks) and which are only used for convenience of calculation. In others they are simply figures adopted as those which in default of definite knowledge appear most probable. They must always be clearly indicated as interpolations; it is always well to state the method by which they are obtained, and any subsidiary information which may be regarded as direct evidence of their accuracy, and if practicable they may be given not as exact, but as lying between certain limits; thus the interpolated figures for Cheshire might be written *12s. 6d. to 13s. 6d.*, instead of *13s. 1d.*

Necessity for distinguishing interpolated figures.

Several different cases are met with in interpolation, some of which are treated algebraically in the next section, while others can be illustrated at once by numerical examples.

THE GRAPHIC METHOD.—If we know the values of quantities at isolated positions, such as the numbers of the population at the ages 25 to 35, 35 to 45, &c.; the population in 1871, 1881, 1891, &c.; wages in 1860, 1870, 1873, &c.; the numbers whose wages are from 15s. to 20s., 20s. to 25s., &c., we may represent the facts by such a diagram as—

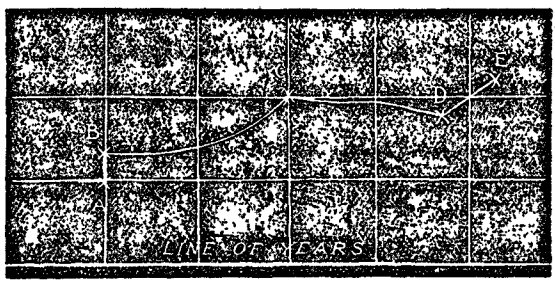
Graphic method.



Suppose that we need the value of the quantity in 1875. If we were only given the two points C and D, the simplest hypothesis, and the one to be made in the absence of any evidence to the contrary, is that the quantity increased uniformly between C and D; representing such an increase by the straight line C D, the height of the point x will represent the quantity in 1875.

If the point E is also given, the hypothesis represented by the straight lines C D, D E will not stand, for it assumes a sudden break in the regularity at the point D in 1877, for which there is no evidence. We must take into account all the points given, and through them all a line must be drawn whose curvature is as smooth as possible, for in the absence of evidence to the contrary, sudden changes in the quantities may be assumed not to exist. Such a curve can be constructed on mathematical principles, or may be drawn freehand; if the latter, it will often be quite as near the facts as the argument will allow us to go.

This method only applies to continuous quantities, such as numbers at different ages, population at different dates, earners at different wages in a very large group of wages. Thus for all England the average wage must change gradually, but the wage of the London builders changed suddenly as the result of strikes and arrangements at certain dates. In this case we must draw the figure to correspond as closely as possible to the evidence, such as—



where A B represents a sudden rise; B C a gradually accelerated increase due to improving trade, C D a slow falling off from the wage reached at C, and D E a determined and successful effort to recover the lost ground.

PERIODIC FIGURES.—If we know the annual averages of figures which have a yearly period and a sufficient number of monthly averages to estimate the periodic fluctuations by the method described on pp. 182-7, we can interpolate figures for any month for which the returns are incomplete with fair accuracy. Thus if we are dealing with the numbers of unemployed as given in the *Labour Gazette*, we find a periodicity which is not very strongly marked in all the months, but there is in general a fall

in the spring and a rise in the late autumn, and June is generally the minimum month. We can then make use of the small diagrams on p. 184, and, having marked in all the information we have, draw the waves on the rising, stationary, or descending line of averages, so that the fluctuating lines shall pass through all the given points. We can obtain an idea of the accuracy of the resulting figures by noticing the general characteristics of the given figures; we find that the percentage unemployed has never changed more than two units in one month, that there are no fluctuations which have lasted less than three or four months, and that the percentages have never been below 1 or above 10. Finally, we can look at the trade history of particular dates, and in the light we thus obtain reject any improbable figures.

USE OF SUBSIDIARY CURVES.—If we are able, by the methods described in Chapter VII., Sect. III., or Sect. V., to find a close connection between two series, we can use the more complete of them to assist the interpolation of any missing figures in the other. We must first investigate carefully the closeness and nature of correspondence at the dates for which we have complete figures in both series. Then we can draw diagrams, similar to those facing p. 175, one of the lines being incomplete. Then completing the broken line, so as to bring it into as close resemblance with the completed line as the given points allow, we shall obtain the most probable values for the missing figures. The accuracy of the result can be tested as in the previous case. This method may reasonably be used in interpolating figures for the yield from one source of revenue by means of the yield from another; for the value of exports from that of imports; for the marriage rate from foreign trade; for the wages in one district from those in another; for the number of unemployed from the changes in consumption of foods; for changes in parts of the population, when we know the changes in the whole, and for many other series.

GENERAL.—Series of figures may be classed in three groups—(1) Periodic; (2) symptomatic, where there is a general tendency towards increase (as in serial wage statistics) or decrease (as in the English birth rate in recent decades); (3) those which have no period and no symptom, but only apparently random fluctuations.

General classification of series.

To interpolate figures in series of the third group it is neces-

sary to obtain a measure of the fluctuations, by the theorems of Part II., and then we can assign the mathematical probability of the various numbers possible. In series of the second group, we must pay attention in addition to the symptomatic tendency. The necessity for interpolation of this kind does not, however, arise frequently, so we will not offer detailed illustrations of it.

SECTION 2.—ALGEBRAIC TREATMENT.

The problem of interpolation to which most attention has been given may be stated as follows:—When one quantity is subject to continuous regular change, and a second quantity changes in connection with it, and we know or can estimate directly only some discontinuous values of this second quantity, it is required to find the probable values of the second quantity which correspond to given values of the first: for instance, given the expectation of life at the ages 15, 20, 25, &c., it is required to find it for intermediate ages; given the population of the country in 1871, 1881, 1891, 1901, find it at intermediate dates. The only permissible assumptions are that the quantity changes continuously, that is with no breaks at any figure, and that the rate of change of the quantity is also continuous, that is that the line representing its value is not angular, but smooth. This problem differs from those just discussed, in that there is likely to be a law binding such figures together, whereas in the former cases the consecution was apparently random.

It is necessary to divide the problem into two classes:—
A. Where the given values may be assumed to be accurate;
B. Where the given values are liable to correction.

A. Some preliminary algebra is necessary; it is derived principally from Boole's *Finite Differences* and De Morgan's *Differential Calculus*, to which authorities readers may be referred for more detailed treatment.

1. Let y be a continuous function of x, and let $y_0, y_1, y_2 \ldots$ be any values of y.

Let $\Delta_0^1, \Delta_1^1, \Delta_2^1 \ldots$ be written for $y_1 - y_0,\ y_2 - y_1,\ y_3 - y_2, \ldots$
$\Delta_0^2, \Delta_1^2, \Delta_2^2 \ldots$ „ $\Delta_1^1 - \Delta_0^1,\ \Delta_2^1 - \Delta_1^1,\ \Delta_3^1 - \Delta_2^1, \ldots$
$\Delta_0^3, \Delta_1^3, \Delta_2^3 \ldots$ „ $\Delta_1^2 - \Delta_0^2,\ \Delta_2^2 - \Delta_1^2,\ \Delta_3^2 - \Delta_2^2, \ldots$
and so on.

$\Delta_0^1, \Delta_1^1, \ldots$ are called differences of the 1st order;
$\Delta_0^2, \Delta_1^2, \ldots$ „ „ 2nd order, and so on.

It is easy to show that $\Delta_0^2 = y_2 - 2y_1 + y_0$, $\Delta_1^2 = y_3 - 2y_2 + y_1$, &c.
$\Delta_0^3 = y_3 - 3y_2 + 3y_1 - y_0$, $\Delta_1^3 = y_4 - 3y_3 + 3y_2 - y_1$, &c.

and generally that—

$$\Delta_o^r = y_r - r.y_{r-1} + \frac{r(r-1)}{1.2}.y_{r-2} - + \text{ to } \overline{r+1} \text{ terms} \quad - \quad - \quad (\alpha)$$

and $$\Delta_s^r = y_{r+s} - r.y_{r+s-1} + \frac{r(r-1)}{1.2}.y_{r+s-2} - + \text{ to } \overline{r+1} \text{ terms} \quad - \quad (\beta)$$

the coefficients being those of the binomial expansion. Equations (α) and (β) are easily proved by induction.

We can also express values of y in terms of y_o and the differences; $y_1 = y_o + \Delta_o^1$, $y_2 = y_1 + \Delta_1^1 = y_o + 2.\Delta_o^1 + \Delta_o^2$, $y_3 = y_o + 3.\Delta_o^1 + 3\Delta_o^2 + \Delta_o^3$, . . . and generally—

$$y_r = y_o + r.\Delta_o^1 + \frac{r(r-1)}{1.2}.\Delta_o^2 + \frac{r(r-1)(r-2)}{1.2.3}.\Delta_o^3 + \ldots + \text{to } \overline{r+1} \text{ terms } (\gamma)$$

$$\Delta_r^s = \Delta_o^s + r.\Delta_o^{s+1} + \frac{r(r-1)}{1.2}.\Delta_o^{s+2} + \frac{r(r-1)(r-2)}{1.2.3}.\Delta_o^{s+3} \ldots + \text{to } \overline{r+1} \text{ terms } (\delta)$$

$$y_{r+s} = y_s + r.\Delta_s^1 + \frac{r(r-1)}{1.2}.\Delta_s^2 + \frac{r(r-1)(r-2)}{1.2.3}.\Delta_s^3 + \ldots + \text{to } \overline{r+1} \text{ terms } (\epsilon)$$

which formulæ can be proved by induction, and in fact (γ) is a special case both of (δ) and (ϵ).

2. If we assume that y can be expanded in ascending powers of x, and is a rational function of the n^{th} order, we have—

$$y = a_o + a_1 x + a_2 x^2 + \ldots a_n x^n, \quad - \quad - \quad - \quad - \quad - \quad - \quad (\zeta)$$

where a_o, a_1, . . . a_n are constants.

Suppose now that y_o, y_1, y_2 . . . y_n are the values of y which correspond to values of x, increasing in arithmetic progression, viz., x_o, $x_o + h$, $x_o + 2h$, . . . $x_o + nh$; then, on this assumption, we have $\overline{n+1}$ equations to determine the constants in equation (ζ).

We can, however, write equation (ζ) in terms of the y's and x's without evaluating the constants; for

$$y = y_o + \frac{x - x_o}{h}.\Delta_o^1 + \frac{x - x_o}{h}.\frac{x - x_o - h}{2h}.\Delta_o^2 + \frac{x - x_o}{h}.\frac{x - x_o - h}{2h}.\frac{x - x_o - 2h}{3h}.\Delta_o^3$$

$$+ \ldots \text{ to } \overline{n+1} \text{ terms} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (\eta)$$

is an equation of the n^{th} degree in x, and reduces to the identity (γ) above, if we substitute $(x_o + rh)$ and y_r, where r is an integer not greater than n, for x and y.

Hence equation (η) is of the same order and satisfied by the same $\overline{n+1}$ pairs of values as equation (ζ), and is therefore identical with it.

Again, if y_o, y_1, . . . y_n are values of y corresponding to *any* values of x, viz., x_o, x_1, . . . x_n, it can be shown that the equation

$$y = y_o.\frac{(x - x_1)(x - x_2) \ldots (x - x_n)}{(x_o - x_1)(x_o - x_2)(x_o - x_n)} + y_1 \frac{(x - x_o)(x - x_2) \ldots (x - x_n)}{(x_1 - x_o)(x_1 - x_2)(x_1 - x_n)} + \ldots$$

$$+ y_n.\frac{(x - x_o)(x - x_1) \ldots (x - x_{n-1})}{(x_n - x_o)(x_n - x_1) \ldots (x_n - x_{n-1})} \quad - \quad - \quad \textit{Lagrange's formula } (\theta)$$

is equivalent to (ζ), for it is of the same degree in x, and is satisfied by the same $\overline{n+1}$ pairs of values of (x, y), viz., (x_0, y_0) $(x_1, y_1) \ldots (x_n, y_n)$.

3. Still assuming that an equation of the form of (ζ) satisfies the conditions, we can at once interpolate any values needed.

Thus, if we are given that $y_0, y_1, y_2, \ldots y_{n-1}$ are values corresponding to $x = 1, 2, 3 \ldots n$ respectively, we can find y_s where s is fractional, by putting $x_0 = 1$, $h = 1$, $x = 1 + s$ in equation (η). We obtain—

$$y_s = y_0 + s.\Delta_0^1 + \frac{s(s-1)}{1.2}.\Delta_0^2 + s\,\frac{(s-1)(s-2)}{1.2.3}.\Delta_0^3 + \text{ to } n \text{ terms} \qquad (\iota)$$

We can easily obtain similar formulæ for any other intervals.

4. Notice that, if $y_0, y_1, \ldots y_n$ correspond to values of x ($x_0, x_1 + h$, &c.) in arithmetic progression—

$$y_0 = a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n,$$

and $y_1 = a_0 + a_1 (x_0 + h) + \ldots + a_n (x_0 + h)^n$ from (ζ);

$$\therefore \Delta_0^1 = a_1 h + \ldots + a_n \{(x_0 + h)^n - x_0^n\}$$

$$= a_1 h + \ldots + a_n . nhx_0^{n-1} + \text{terms of lower degree in } x,$$

an equation of the $\overline{n-1}^{\text{th}}$ degree only.

Continuing this process, we obtain—

$$\Delta_0^n = a_n . h^n . n!, \text{ and there are no higher differences.}$$

Also, since (y_0, x_0) $(y_1, x_0 + h) \ldots \left(y_{n+1}, x_0 + \overline{n+1}.h\right)$ lie in a curve of the n^{th} degree, we have from equation (α)—

$$y_{n+1} - (n+1).\ y_n + \frac{(n+1).n}{1.2}.\ y_{n-1} - \frac{(n+1).n.(n-1)}{1.2.3}.\ y_{n-2}$$

$$+ - \text{ to } \overline{n+2} \text{ terms} = \Delta_0^{n+1} = 0 \qquad (\kappa)$$

5. If for any purpose we need to evaluate the constants in equation (ζ), we can abbreviate the solution, as follows, if the x's are in arithmetic progression.

Given five pairs of values, we have—

$$y_0 = a_0 + a_1 x_0 + a_2 x_0^2 + a_3 x_0^3 + a_4 x_0^4$$

$$y_1 = a_0 + a_1 (x_0 + h) + a_2 (x_0 + h)^2 + a_3 (x_0 + h)^3 + a_4 (x_0 + h)^4$$

$$y_2 = a_0 + a_1 (x_0 + 2h) + a_2 (x_0 + 2h)^2 + a_3 (x_0 + 2h)^3 + a_4 (x_0 + 2h)^4$$

$$y_3 = a_0 + a_1 (x_0 + 3h) + a_2 (x_0 + 3h)^2 + a_3 (x_0 + 3h)^3 + a_4 (x_0 + 3h)^4$$

$$y_4 = a_0 + a_1 (x_0 + 4h) + a_2 (x_0 + 4h)^2 + a_3 (x_0 + 4h)^4 + a_4 (x_0 + 4h)^4$$

As in the last paragraph $\Delta_0^4 = a_4 . h^4 . 4!$, $\therefore\ a_4 = \dfrac{\Delta_0^4}{24h^4}$ - - - (λ)

It is easily seen that Δ_0^3 is independent of a_0, a_1 and a_2, and that

$$\Delta_0^3 = a_3 \{(x_0 + 3h)^3 - 3(x_0 + 2h)^3 + 3(x_0 + h)^3 - x_0^3\} +$$

$$a_4 \{(x_0 + 3h)^4 - 3(x_0 + 2h)^4 + 3(x_0 + h)^4 - x_0^4\} = 6h^3 a_3 + a_4 \{24h^3 x_0 + 36h^4\}$$

whence $a_3 = \dfrac{\Delta_0^3}{6h^3} - \Delta_0^4 \left(\dfrac{x_0}{6h^4} + \dfrac{1}{4h^3}\right)$;

while $\Delta_o^2 = 2h^2.a_2 + a_3(6h^2x_o + 6h^3) + a_4(12h^2x_o^2 + 24h^3x_o + 14h^4)$, which gives a_2. a_o and a_1 can then be found from the first two equations. The points of inflexion on the curve, $y = a_o + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, are determined by the equation—

$$o = \frac{d^2y}{dx^2} = 2a_2 + 6a_3x + 12a_4x^2$$

and the sign of $\frac{d^3y}{dx^3}$, *i.e.* of $a_3 + 4a_4x$, decides the nature of the change of curvature.

This method is employed on page 254 *infra*.

6. In evaluating the constants it will be found that the following identities are sometimes useful :—

$n^r - {}^sC_1\overline{n-1}^r + {}^sC_2\overline{n-2}^r - + =$ the coefficient of x^r in the expansion of $r!\{e^{nx} - {}^sC_1e^{\overline{n-1}x} + {}^sC_2e^{\overline{n-2}x} - \}$

i.e. of $r!e^{\overline{n-s}x}(e^x - 1)^s$

i.e. of $r!\left(1 + \overline{n-s}x + \frac{\overline{n-s}^2}{2}x^2 + \right).\ x^s\left(1 + \frac{x}{2} + \frac{x^2}{2.3} + \right)^s.$

Coefficient is o, when $r = 1, 2, 3 \ldots s-1$.

„ $r!$, when $r = s$.

„ $\left(n + \frac{s}{2}\right).r!$, when $r = s+1$. - - (μ)

7. It is necessary to express the differential coefficients of y with regard to x in equation (ζ) in terms of the differences, and conversely. Now from a comparison of (ξ) and (η) we have, when x_o is zero—

$$y = a_o + \frac{x}{h}(\Delta_o^1 - \tfrac{1}{2}\Delta_o^2 + \tfrac{1}{3}\Delta_o^3 - \tfrac{1}{4}\Delta_o^4 +) + a_2x^2 + a_3x^3 + +$$

$$\therefore\ h\frac{dy}{dx} = (\Delta_o^1 - \tfrac{1}{2}\Delta_o^2 + \tfrac{1}{3}\Delta_o^3 - \tfrac{1}{4}\Delta_o^4 +) + x(\quad).$$

Writing $y = f(x)$ for equation (ζ), the equation just written gives, when $x = x_o = o$,

$$hf^1(o) = \Delta_o^1 - \tfrac{1}{2}\Delta_o^2 + \tfrac{1}{3}\Delta_o^3 - \tfrac{1}{4}\Delta_o^4 +$$

Applying the same process again and again, and remembering that Δ_o^2 bears the same relation to Δ_o^1 as Δ_o^1 bears to y_o, and so on, we obtain, omitting the suffixes—

$$h^nf^n(o) = (\Delta - \tfrac{1}{2}\Delta^2 + \tfrac{1}{3}\Delta^3 -)^n y_o \qquad (\nu)$$

where the Δ's are to be treated as ordinary algebraic quantities, till the exponent is removed.

Thus $h^2f^2(o) = \Delta^2(1 - \frac{1}{2}\Delta + \frac{1}{3}\Delta^2 -)^2$

$= \Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 -$

$h^3f^3(o) = \Delta^3 - \frac{3}{2}\Delta^4 +$ &c.

It is to be noticed in particular that each derived function depends only on differences of as high an order as itself.

Again, by Taylor's Theorem,

$$y_1 = f(x_o + h) = f(x_o) + hf^1(x_o) + \frac{h^2}{2!}f^2(x_o) +$$

$$y_o = f(x_o)$$

$$\therefore \Delta_o^1 = y_1 - y_o = hf^1(x_o) + \frac{h^2}{2!}f^2(x_o) +$$

$$y_2 = f(x_o + 2h) = f(x_o) + 2hf^1(x_o) + \frac{4h^2}{2!}f^2(x_o) +$$

$$\therefore \Delta_o^2 = y_2 - 2y_1 + y_o = h^2 f^2(x_o) + h^3\{\qquad\}.$$

and, using equations (α) and (μ), or otherwise, generally

$$\Delta_o^s = h^s f^s(x_o) + h^{s+1}\{\qquad\} \qquad (o)$$

8. We may now consider the assumptions made when we took (ζ) to express the relation between y and x.

If y and x are connected by any functional law, that is if y is determinate for all given values of x, without which assumption interpolation is meaningless, then y can be expressed as a function of x; let $y = f(x)$, then, by Maclaurin's theorem—

$$y = f(x) = f(o) + xf^1(o) + \frac{x^2}{2}f^2(o) + \frac{x^3}{3}f^3(o) + +$$

to an infinite number of terms.

If $f^{n+1}(o)$ and following coefficients are very small, and x is never large, the terms from the $\overline{n+2}^{th}$ onwards become negligible in comparison with earlier terms, so that the first $\overline{n+1}$ terms determine the value of y approximately. Now by the equations (ν) and (o), f^{n+1} is small when Δ^{n+1}, Δ^{n+2}, . . . are small, and *vice versa*. Hence we have the following general statement: any functional relation between y and x reduces to the parabolic equation of the n^{th} degree (ζ), if the differences of orders higher than the n^{th} vanish, and if these differences do not vanish but are small, equation (ζ) is still an approximate expression for the relation.

Now if the line drawn through the given points is to have continuous and slowly changing curvature, it is easily verified that the second differences for points near together are not large, for a rapid change in the rate of increase of the ordinate means a rapid change of curvature; and if we construct a second curve with the same abscissæ and the first differences as ordinates,

small third differences will indicate absence of rapid change in the first, and so on; but beyond this point it is not easy to see the connection between the hypothesis underlying interpolation and the diminution of successive differences. The converse, however, is clearer; if in any series of figures it is found experimentally that the successive differences tend to disappear, then any curve which passes through the points is expressed approximately by the parabolic equation. De Morgan states this conclusion thus:—"If we take n points near each other, and having their abscissæ in arithmetic progression, with a small or at least not very large common difference, and their ordinates not very unequal . . . the parabola of the $\overline{n-1}^{th}$ order will very nearly coincide with any regular curve of the same general appearance, at least between the same points." Boole's explanation is:—"It is customary to assume for the general expression of the values under consideration a rational and integral function of x, and to determine the constants by the given conditions. This assumption rests upon the supposition (a supposition, however, actually verified in the case of all tabulated functions *) that the successive orders of differences rapidly diminish."

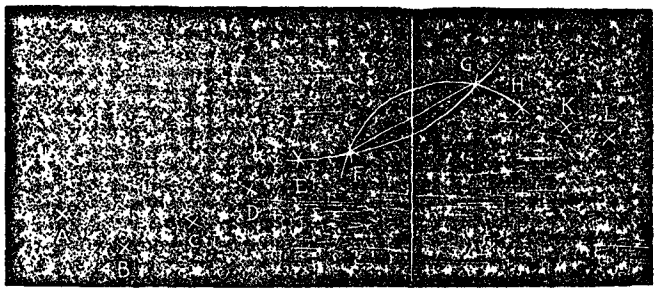
Since, from equation (o), when h is small, the successive differences for any curve diminish as their order becomes higher, it is a legitimate process to build up a series of values of any function on the hypothesis that the higher differences vanish.

If a freehand curve is drawn so as to pass through the chosen fixed points, and to have curvature which changes as slowly as possible, a line will be obtained which lies very near that given by equation (ζ). Such a line would be similar to the track of a bicyclist who was riding so as to pass over several marks, or to just avoid several obstacles.

9. It is clear from the above analysis that we can make a smooth continuous curve pass through any number of points we please; for with the parabolic equation (ζ) there are never any sudden jumps in the values of y, $\frac{dy}{dx}$, or $\frac{d^2y}{dx^2}$, as x changes continuously; and we can obtain as many linear equations (which have always real values) as there are constants, simply by taking n in the original equation to be the number of fixed points.

* That is mathematical functions such as $\int_0^x e^{x-x^2}\,dx$, not statistical approximations.

If we have, let us say, 10 points, as—



and wish to find a point on a fixed vertical line between F and G, we can either take only F and G into consideration, and, joining them by a straight line, obtain the point x_1; or considering E, F, and G, or F, G, and H, draw parabolas and obtain x_2 or x_3; or considering E, F, G, and H, draw a parabola of the third order, which would have a point of inflexion near F; this would be approximately the path a bicyclist might follow if he had to start from E, and ride to a near point H, passing close to F and G. If we now include D and K (if our bicyclist has to start from D, pass E, F, G, and H, and reach K) we shall modify the curvature throughout; and as we include more and more points shall continue to affect slightly the path F G. If the inclusion of the nearer points tends to make the line F G approximate more and more closely to a final position, while the further inclusion of the more distant points throws it further away, we may conclude that the positions of these further points are not governed by the same numerical conditions as the nearer one. Thus in a "table of survivals" the figures for ages under 5 years are not distributed in accordance with the curve determined by the figures for higher ages; in a table showing wages, it may be seen that those of highly paid workmen are not governed by the same causes as those lower in the scale. On the other hand, the number in each census is dependent on all the previous numbers for more than one generation. In interpolating for the population of 1876 we shall obtain different figures according as we include 1851, '61, '71, '81, '91 only, or 1901 as well; and this is not surprising, for a mistake made in 1876 may not come to light till we have watched the growth of the population for twenty-five years. It is clear that the points far from the period in which the interpolation is to be done cannot be allowed so much influence as

those nearer, and it appears experimentally that this condition is fulfilled in the method discussed; also, in series (η) the successive coefficients begin to diminish with the r^{th} term where $x < x_0 + (2r-3)h$, that is with the coefficient of the first difference when x is between x_0 and $x_0 + h$. It may be noticed that the wanderings of the curve are limited by the condition that a curve of the $n-1^{th}$ order cannot have more than $\overline{n-3}$ points of inflexion, for $\frac{d^2y}{dx^2}$ has no term of a higher degree than x^{n-3}.

In the above illustration the intermediate points from F to G might be found from the five points D, E, F, G, H, or from E, F, G, H, K. These two curves may be welded together between F and G. The points near F are more accurately determined by the first, of which it is the middle; those near G by the second. The welding line should touch the first at F, the second at G. This is conveniently done by the use of the sine curve. This method is employed, I believe, at the Registrar-General's office.

It cannot be said that the present theory of statistical interpolation rests on an altogether satisfactory basis.* The principles which govern it are not well defined, and the mathematical analysis of the methods, by which the principles should be brought into relation with the facts, is incomplete. Yet it is perhaps unnecessary to labour after more refined methods, for interpolation cannot be precise unless we actually know the algebraic expression of the laws which govern the figures, and the method here discussed is found to satisfy the conditions empirically, while further refinements could only introduce slight modifications.

* This remark does not apply to the interpolation in evaluating mathematical functions.

10. EXAMPLES SHOWING THE NUMERICAL USE OF THE FORMULÆ.—(1.) Given the number of wage-earners earning sums in 5s. groups, to estimate the number earning as much as 24s. and not so much as 25s.

		*Numbers per 1,000 Wage-Earners (Adult males)	DIFFERENCES.			
			1st.	2nd.	3rd.	4th.
Earning as much as 10s. but not so much as	15s.	39	257			
	20s.	296		46		
	25s.	599	303		−144	
	30s.	804	205	−98		151
	35s.	918	114	−91	7	18
	40s.	966	48	−66	25	

Neglect the increasing differences arising from the number earning less than 15s.

Using formula (η), $x_0 = 20$ (shillings), $h = 5$, $y_0 = 296$, $\Delta_0^1 = 303$, $\Delta_0^2 = -98$, $\Delta_0^3 = 7$, $\Delta_0^4 = 18$.

At 25s., $y = 599$, from above table.

$$\text{At 24s., } x = 24,\ y = 296 + \frac{4}{5} \text{ of } 303 + \frac{4}{5}\cdot\frac{-1}{10} \text{ of } (-98) +$$

$$\frac{4}{5}\cdot\frac{-1}{10}\cdot\frac{-6}{15} \text{ of } 7 + \frac{4}{5}\cdot\frac{-1}{10}\cdot\frac{-6}{15}\cdot\frac{-11}{20} \text{ of } 18.$$

$$= 296 + 242.4 + 7.84 + .224 + .3168 = 547 \text{ (nearly).}$$

The required number is therefore $599 - 547 = 52$.

Again at 23s., $x = x_0 + 3$, $y = 489$, and the number earning as much as 23s. and not so much as 24s. is 58.

(2.) To make an estimate for the value of imports in the year 1813, the records for which were destroyed by fire.

Given value of imports in—

1810	£39,202,000	y_1.
1811	26,510,000	y_2.
1812	26,163,000	y_3.
1813	...	y_4.
1814	33,755,000	y_5.
1815	32,987,000	y_6.
1816	27,431,000	y_7.

* General Report on Wages.

From formulæ (κ), using y_3 and y_5 only—

$$y_5 - 2y_4 + y_3 = o, y_4 = 29{,}959.$$

From formulæ (κ), using y_2 and y_6 as well—

$$y_6 + y_2 - 4(y_5+y_3) + 6y_4 = o, y_4 = 30{,}029.$$

From formulæ (κ), using y_1 and y_7 as well—

$$y_7 + y_1 - 6(y_6 + y_2) + 15(y_5+y_3) - 20y_4 = o, y_4 = 30{,}421.$$

Here the first and second values are very near together, while the third differs; hence we adopt £30,000,000 as the value required. (*Cf.* a similar example in Boole's chapter.)

(3.) In Mr Booth's *Life and Labour of the People, e.g.*, Vol. V., p. 46, a series of very useful diagrams is given showing the age distribution of various classes. The figures he uses are as follows:—

Ages.	Proportion occupied per 10,000 of total aged 10-80.	Average at each year of age between given limits.
10-15 years	193·5	38·7
15-20 „	880	176
20-25 „	933	188·6
25-35 „	1,636	163·6
35-45 „	1,201	120·1
45-55 „	830	83
55-65 „	434	43·4
65-85 „	192·5	12·8

His diagram is drawn from the last column, the numbers in which form the ordinates for the middle of the corresponding age periods. The points so obtained are joined by straight lines. This method is sufficiently accurate for his purpose, but it will afford an interesting example of interpolation if we obtain some of the figures for intermediate years more closely.

Mean Age.	Numbers up to corresponding limits.
$x_1 = 12\frac{1}{2}$	$y_1 = 193{\cdot}5$
$x_2 = 17\frac{1}{2}$	$y_2 = 1073{\cdot}5$
$x_3 = 22\frac{1}{2}$	$y_3 = 2006{\cdot}5$
$x_4 = 30$	$y_4 = 3642{\cdot}5$
$x_5 = 40$	$y_5 = 4843{\cdot}5$
$x_6 = 50$	$y_6 = 5673{\cdot}5$
$x_7 = 60$	$y_7 = 6107{\cdot}5$
$x_8 = 72\frac{1}{2}$	$y_8 = 6308$

Since the x's are not in arithmetic progression, we must use Lagrange's formula (θ).

To find the ordinate corresponding to the age 35, for example, we will include the five values of y from y_2 to y_6.

$$\text{Then } y = 1073.5 \times \frac{12\tfrac{1}{2}.5.(-5)(-15)}{(-5)(-12\tfrac{1}{2})(-22\tfrac{1}{2})(-32\tfrac{1}{2})} + 2006.5 \times \frac{17\tfrac{1}{2}.5.(-5).(-15)}{5.(-7\tfrac{1}{2})(-17\tfrac{1}{2})(-27\tfrac{1}{2})}$$
$$+ 3642.5 \times \frac{17\tfrac{1}{2}.12\tfrac{1}{2}.(-5)(-15)}{12\tfrac{1}{2}.7\tfrac{1}{2}.(-10)(-20)} + 4843.5 \times \frac{17\tfrac{1}{2}(12\tfrac{1}{2})(5)(-15)}{22\tfrac{1}{2}.(17\tfrac{1}{2}).10.(-10)}$$
$$+ 5673.5 \times \frac{17\tfrac{1}{2}.12\tfrac{1}{2}.5.(-5)}{32\tfrac{1}{2}.27\tfrac{1}{2}.20.10}$$
$$= 4412.$$

Mr Booth's method gives 4243 for the same position.

(4.) We can now determine the median and the mode more accurately than before. We will use the figures already employed in Chapter IV., which may be retabulated thus:—

Earning more than	Numbers.	Differences.				Corresponding Abscissæ.
$4.75	9					19
4.25	13					17
3.75	109					15
3.25	363					13
2.75	561	506				11
2.25	1,067	970	464	−137		9
1.75	2,037	1,297	327	−152	−15	7
1.25	3,334	1,472	175	−1330	−1178	5
.75	4,806	317	−1155			3
.25	5,123					1

Unit of abscissa, $.25.

To find the median use the five points whose abscissæ are 11, 9, 7, 5, 3.

Equation (ξ) gives—

$$561 = a_0 + a_1.11 + a_2.11^2 + a_3.11^3 + a_4.11^4$$
$$1067 = a_0 + a_1.9 + a_2.9^2 + a_3.9^3 + a_4.9^4.$$
$$2037 = a_0 + a_1.7 + a_2.7^2 + a_3.7^3 + a_4.7^4$$
$$3334 = a_0 + a_1.5 + a_2.5^2 + a_3.5^3 + a_4.5^4$$
$$4806 = a_0 + a_1.3 + a_2.3^2 + a_3.3^3 + a_4.3^4$$

Using equations (λ), we have—

$$a_4 = \frac{-15}{24 \times 2^4} \mathrel{\mathfrak{C}} -\frac{5}{128}, \text{ since } h = -2,\ \Delta_0^4 = -15;\ \mathfrak{C}$$

$$a_3 = \frac{137}{6 \times 8} + 15\left(\frac{11}{6 \times 16} - \frac{1}{4 \times 8}\right), \text{ since } x_0 = 11 \text{ and } \Delta_0^3 = \cdot 137,$$
$$\mathfrak{C} = \frac{197}{48}.$$

$$464 = 8a_2 + \frac{197}{48}(24 \times 11 - 6 \times 8) - \frac{5}{128}(48 \times 11^2 - 24 \times 8 \times 11 + 14 \times 16)$$

$$a_2 = -33\tfrac{43}{64}.$$

Equations (μ) could have been used with advantage if the difference between successive abscissæ had been unity.

a_1 is found from the equation—

$$-1472 = 2a_1 + 16a_2 + 98a_3 + 544a_4$$

$$a_1 = -657\tfrac{5}{48}$$

and finally $a_0 = 6972\tfrac{91}{128}$.

The median is then found from the equation—

$$2561\tfrac{1}{2} = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$$

Solving by Horner's method, we find $x = 6.142$; and, therefore, the median is at \$1.536.

Second method :—

Suppose x expressed as a function of y* and apply Lagrange's formula (θ) suitably altered.

$$x = 3.\frac{2)561\tfrac{1}{2} - 3334)(2561\tfrac{1}{2} - 2037)(2561\tfrac{1}{2} - 1067)(2561\tfrac{1}{2} - 561)}{(4806 - 3334)(4806 - 2037)(4806 - 1067)(4806 - 561)}$$

$$+5.\frac{(2561\tfrac{1}{2} - 4806)(\quad)(\quad)}{(3334 - 4806)(\quad)(\quad)} \quad +7\times \quad +9\times \quad +11\times,$$

whence $x = 6.237$; that is the median is at \$1.56.

This method saves the solution of a biquadratic, and with small numbers would need less numerical work than the first method.

Third method :—

Use formula (η) to obtain the necessary equation.

$$\text{Thus } 2561\tfrac{1}{2} = y = 561 + \frac{x-11}{-2} \text{ of } 506 + \frac{(x-11)(x-9)}{-2\times-4} \text{ of } 464$$

$$+ \frac{(x-11)(x-9)(x-7)}{-2\times-4\times-6} \text{ of } (-137)$$

$$+ \frac{(x-11)(x-9)(x-7)(x-5)}{-2\times-4\times-6\times-8} \text{ of } (-15).$$

This reduces to the same equation as in the first method—

$$2561\tfrac{1}{2} = 6972\tfrac{91}{128} - 657\tfrac{5}{48}x - 33\tfrac{43}{64}x^2 + \frac{197x^3}{48} - \frac{5x^4}{128}$$

The quartiles, deciles, and percentiles can be found by similar methods.

---

* Compare Edgeworth's *Representation of Statistics by Mathematical Formula, Statistical Journal*, 1898, p. 699.

For the mode we must take in the last number, 5123, and recalculate a_2, a_3, a_4 for the five highest values of y, and then solve the quadratic given in paragraph 5, viz.—

$$6a_4x^2+3a_3x+a_2=0,$$

giving the constants their new values.

Hence $x = 8.2$ or 4.40; $\frac{d^3y}{dx^3}$ is positive and $\therefore \frac{d(-y)}{dx}$ a maximum, when $x=4.40$. The mode is then at \$1.10.

The mode can of course be determined less accurately by taking 4 or 3 given points instead of 5, or for greater accuracy more can sometimes be used.

Another mode may be found between \$2.75 and \$3.75 from the five highest abscissæ. This proves to be at \$3.20.

This method is applicable to such problems as the determination of the date at which the population, the marriage, birth, and death rates, &c., increased most rapidly; at what age the chance of death increases most, &c.*

B. The second division of the problem of interpolation is when the original returns have to be corrected, *e.g.*, the determination of the distribution by age from the census returns.

We have now the problem of drawing a smooth line in the neighbourhood of a great number of points, but not necessarily through any of them. The assumption is that the returns are insufficient in number or deficient in accuracy, and that they indicate a regular distribution which it is required to represent.

1. One method is to assume that the averages over fairly large groups are accurate, and to these averages to apply any of the methods discussed under group A.

2. A second method has been used in the section in which various curves were smoothed (*vide supra*, Chapter VII.). This may be restated as follows:—Take successive groups of 2, or 3, or 4 . . . . 10 points, beginning again and again at the ordinates for each of the given abscissæ. Find the centres of gravity of each group; that is, erect an ordinate equal to the average of the ordinates of a group at the point half-way between the ends of the abscissæ of the outside ordinates of the group. Draw a line through the points so obtained. It will be found that this line satisfies all the conditions laid down. An example of this method is given in the diagram facing p. 151.

* *Cf.* Edgeworth, in *Statistical Journal*, 1899, p. 381, and the references there given.

3. In another method* the original figures are smoothed till the differences of the fourth or fifth or higher orders vanish; and then the ordinary formulæ of interpolation are applied.

Thus in example 1, on page 250, rewrite the table thus:—

Wages.	Smoothed Numbers.	Corrected Differences.		
		1st.	2nd.	3rd.
Up to 20s.	296	$303+a$		
,, 25s.	$599+a$	$205+b$	$-98-a+b$	$7-3b$
,, 30s.	$804+a+b$	$114-a-b$	$-91-a-2b$	$25+2a+3b$
,, 35s.	918	48	$-66+a+b$	
,, 40s.	966			

If we put $b=2\frac{1}{3}$, $a=-16$, the third differences vanish, and we have $\Delta_0^1=287$, $\Delta_0^2=-79\frac{2}{3}$, $\Delta_0^3=\Delta_0^4=0$; when $x=25$, $y=583$, and when

$$x=24,\ y=296+\tfrac{4}{5} \text{ of } 287-\tfrac{2}{25} \text{ of } (-79\tfrac{2}{3})=532;$$

so that the number earning as much as 24s. and not so much as 25s. is now found to be 51, instead of 52.

The corrections may be applied to any of the original figures.

We need to solve only one more equation to complete our table from 20s. to 30s.

When $x=23$, $y=296+\frac{3}{5}$ of $287+\frac{3}{25}$ of $79\frac{2}{3}$. The difference between this and the value of y, when $x=24$, is $\frac{1}{5}$ of $287-\frac{1}{25}$ of $79\frac{2}{3}=54.2$.

We have therefore the following table, where the figures in italics have already been calculated, while the others are added on the assumption that the third differences are zero.

Wages.	Numbers.	Differences.		
		1st.	2nd.	3rd.
Up to 20s.	*296*	63.8	...	...
,, 21s.	360	60.6	3.2	0
,, 22s.	420	57.4	3.2	0
,, 23s.	478	*54.2*	3.2	0
,, 24s.	532	*51*	3.2	0
,, 25s.	*583*	47.8	3.2	0
,, 26s.	631	44.6	3.2	0
,, 27s.	677	41.4	3.2	0
,, 28s.	719	38.2	3.2	0
,, 29s.	757	35	3.2	...
,, 30s.	792		...	

If we had taken the second differences more exactly, we

* Suggested to me by Mr W. F. Sheppard.

should have obtained $804 + a + b = 790\frac{1}{3}$ for the last figure as in the previous table.

This method of writing down many figures when the significant differences have been found can be very generally applied in Group A as well as here.

4. Another method, involving higher mathematics, would be discussed more suitably after the section devoted to the law of error; a brief explanation with a useful formula may, however, be offered here.

Suppose we have five consecutive points $(-2, y_2)$, $(-1, y_1)$, (o, y), $(1, y_1)$, $(2, y_2)$ given.

A parabola of the fourth order could be drawn through these five points, but would have two points of inflexion. A great number of parabolas of the third order can be drawn near all the points, having no points of inflexion, and satisfying all the ordinary conditions of interpolation.

Borrowing a principle from the method of least squares,* if the coefficients of the parabola $y = a + bx + cx^2 + dx^3$ are chosen so as to make the quantity

$$\Sigma(a + bx + cx^2 + dx^3 - y)$$

(where the summation extends over the five pairs of values of x and y) a minimum, the parabola so determined will be the best for the purpose.

For the necessary mathematical analysis, Professor Darwin's paper *On Fallible Measures*,† from which this method is taken, should be consulted.

The following equation is obtained—

$a = y - \frac{3}{35} \times \Delta^4_o$, where Δ^4_o is the difference of the fourth order for the y's.

Now replace the point (o, y) by the intersection of its ordinate with the parabola, that is by (o, a), where a has the value just given, that is, diminish y by the quantity $\frac{3}{35}.\Delta^4_o$.

Repeat the same process for each point on the original line, taking it as the middle of a group of 5, and a smooth curve lying very near all the original points is obtained.

---

* See Merriman's *Method of Least Squares*, Chap. III.

† See *Phil. Mag. and Journal*, July 1877.

Thus we may smooth line C in diagram facing p. 164.

Imported Wheat per head of the Population.	Differences.				Smoothed Figures.
lbs.					
1890 226					
	18				
1891 244		− 17			
	1		19	− 16	$245+\frac{3}{35}$ of $16=246\frac{1}{3}$
1892 245		2			
	3		3	13	$248-\frac{3}{35}$ of $13=247$
1893 248		5	16		
	8			− 94	$256+\frac{3}{35}$ of $94=264$
1894 256		21	− 78		
	29			134	$285-\frac{3}{35}$ of $134=263\frac{1}{2}$
1895 285		− 57	56		
	− 28			− 16	$257+\frac{3}{35}$ of $16=258\frac{1}{2}$
1896 257		− 1	40		
	− 29				
1897 228		39			
	+ 10				
1898 238					

5. In many series of observations it is found that the numbers very nearly satisfy some algebraic formulæ,* such as the binomial expansion, the geometric progression, the law of error, or some specially chosen expression. In such a case the constants of the equation chosen are computed by methods similar to that of the last paragraph, and the original observations are replaced by the ordinates of the curve thus determined. Prof. Pareto has found an equation which fits the data of the distribution of incomes.† Modern mathematical statistics deals very frequently in such formulæ. Here we will briefly describe one which has very practical utility, namely, Makeham's formula for the life table.‡ If l_x is the number who survive to the age x out of a given generation, then the formula $l_x = ks^x (g)^{c^x}$, where k, s, g, c are selected constants, fits the records from the ages of 20 upwards with such exactness, that the formula is used for practical actuarial calculations. The formula is not quite arbitrary, but can be obtained from the hypotheses described in the following paragraph.

Let the quantity $\int_{t=0}^{t} \frac{l_{x+t} - l_x}{l_x} = \frac{-dl_x}{l_x} = \mu_x dx$. Then μ_x is called the "force of mortality," and represents the ratio of the number of persons dying in a short interval to the total number alive at the beginning of the interval.

* See Edgeworth, *ibid.*, p. 671.

† See *Statistical Journal*, 1896, p. 533.

‡ See *Institute of Actuaries Text-Book*, Part II.

This force is supposed to consist of two parts, one constant and = A; and the other such that the ratio of the increase of the force to the force is constant, that is, that the force continually increases in a geometric progression. For the latter part (μ'_x)

$$\frac{d\mu'_x}{\mu'_x} = D$$

$$\log \mu'_x = Dx + E$$

$$\mu'_x = e^{Dx+E} = B\,c^x, \text{ where } B = e^E, \text{ and } \log_e c = D.$$

Then $\mu_x = A + B\,c^x$.

This equation represents the hypothesis that the chance of death consists of one part which is constant for all ages, and another which is due to the power of resisting death diminishing continuously with age in a constant ratio.

We have $-\frac{dl_x}{l_x} = \mu_x dx = (A + Bc^x)dx$

$$-\log l_x = Ax + k_1 c^x + k_2 \text{ where } k_1, k_2 \text{ are constants.}$$

$$l_x = k.s^x.(g)^{c^x}, \text{ where } -A = \log_e s$$

$$-k_1 = \log_e g$$

$$-k_2 = \log_e k$$

For further information on the subject of interpolation, the reader is referred to Dr Farr's *Life Table* (No. 3), 1864, Boole's *Finite Differences*, *Text-Book of Institute of Actuaries*, Part II., p. 420 *seq.*, Rice's *Theory and Practice of Interpolation*, 1899, Merrifield *On Quadratures and Interpolation* (British Association Report, 1880), Chauvenet's *Spherical and Practical Astronomy* (Chap. II.), Woolhouse in the *Assurance Magazine* (Vols. XI., XII.), Professor J. D. Everett's Papers (published or forthcoming) *On the Algebra of Difference Tables* (Quarterly Journal of Mathematics, No. 124, 1900), *On a Central-difference Interpolation Formula* (British Association Report, 1900), and in the Journal of the Institute of Actuaries, January 1901, and Mr W. F. Sheppard's Papers *On Central Difference Formulæ* (Proceedings of the London Mathematical Society, Vol. XXXI., Nos. 707-710), and *On the Use of Auxiliary Curves in Statistics of Continuous Variation* (Statistical Journal, September 1900). In these other references will be found. Part of the foregoing chapter might be simplified by the use of "central differences," but in so short an introduction to the subject it seemed best to keep to the more familiar method.

PART II.

APPLICATION OF THE THEORY OF PROBABILITY TO STATISTICS.

PART II.

APPLICATION OF THE THEORY OF PROBABILITY TO STATISTICS.

SECTION I.

INTRODUCTORY.

THE arguments on which the theory of algebraic probability depends are not difficult to follow, and are in fact grounded on every-day experience; the development of calculations also is often little more than straightforward arithmetic; and without using any elaborate mathematical theories we can examine the nature and deduce the equation of the curve of error, which, though it is the foundation of modern mathematical statistics, is only a reasonable summary of common experience.

Object of Part II.

It is not proposed here to go beyond the more elementary and common applications of the law of error; the more advanced treatment tends to deal more with theory and less with practical applications, and is most suitably studied in the original treatises scattered through the journals of various learned societies. The present object is to endeavour to make clear the groundwork of the subject, so that it will be the easier for students to follow modern writers on statistics; the mathematicians who are opening up new ground in this direction naturally cannot stop in each article they write to establish the elementary theorems which are already common property, and so it is often not easy for readers, unfamiliar with these elements, to find any satisfactory discussion or proof of the preliminary formulæ or theorems, since they are not contained in any text-book devoted to the subject. It is this lack of

a preliminary text-book that it is wished to supply in this and the following sections.

The treatment is not intended to be original, and is, it is hoped, not inconsistent with Professor Edgeworth's published treatises,* since the greater part of the mathematics employed is gleaned from his essays, and the earlier authorities to whom he makes reference. The exact form of the proofs employed, and the particular ways in which the formulæ are used, are not in all cases to be found elsewhere; and any fault which may be found with the arguments or application of formulæ must attach to the present writer. To avoid mere repetition of what is better said elsewhere, and not to cumber the ground with well-known elementary formulæ, the reader is assumed to be acquainted with Dr Venn's *Logic of Chance*, and Whitworth's *Choice and Chance*, or with the chapter on Probability to be found in ordinary school algebras. It is hoped that the following pages, however, will be for the most part independent of proofs or formulæ of which the explanation is not furnished.

Need for application of algebraic theory

To the statistician of a generation ago, to the so-called practical men of the present day, and perhaps to some political economists, it would seem absurd and unnecessary to apply these tedious arguments and complicated formulæ to the study of mere figures, which at first sight appear subject to the ordinary rules of arithmetic; but it will be found as we proceed that we are able by their use to solve problems and investigate causal relations which, though apparently simple, must entirely baffle direct attempts to obtain an easy solution. The necessity of some application of the rules of probability becomes evident from the very definition of the science of statistics.†

arises from the definition of statistics.

Statistics deals with great numbers, the numbers of the items which compose some part of the economic or social body as a whole. It does not deal with a single homogeneous mass but with a complex body composed of multitudinous units differing in form and action one from the other; and it is with the complex not with the units that it is concerned. Just as in the mechanics of rigid bodies it is necessary to

* See *Lit. and Phil. Magazine, passim; Statistical Journal, passim;* Report of British Association on *Methods of Ascertaining Variations in the Monetary Standard*, 1888, and others.

† See *supra*, p. 7.

make some hypothesis as to the laws which hold their constituent molecules in place, before any general problems relating to their motion as a whole can be attacked, and in the kinetic theory of gases a generalized theorem of the motion of the separate molecules is employed, so in statistics we must obtain some generalizing principle as to the relation of unit to unit before we can study the phenomena manifested by the body. The economist and the politician, when investigating the effect of a given force, are as a rule concerned with its effect on the whole mass, not on the individuals in particular.* For illustration, we may take one of the numerical totals, relating to a nation, that remains nearly stationary year by year; say the number of marriages yearly in a population of ten millions. It is on the *total* that we trace the effects of a change in our marriage laws. If we regarded only a single family, or a village or small town, we should not find any constancy; the marriage rate would be changing continually with the personnel and age of the small community, and we could not trace with certainty the effect of any external cause. But add family to family, village to village, and district to district; the individual peculiarities of the parts are rapidly lost in the total; in a large community the same number are of marriageable age year by year, the same distribution by age and sex recurs continuously; if undisturbed by external influences, the same marriage rate will be found over a long period. Each couple is influenced by many circumstances before finally deciding to marry; there are very many causes, each of limited effect, which influence the question in different localities, such as an exodus of young men from one district, commercial depression in another, a new demand for labour in a third; but when many districts are taken together these small disturbances counterbalance one another. To produce a change in the rate, the action of a cause is necessary which affects many districts in the same way. Here is to be found the assumption that underlies all statistical investigation, viz., that many independent disturbing causes of small individual effect neutralise one another in the mass. †

It is a matter of common experience that great numbers

* See Mill's *Logic*, Book III., Chap. 23, and Book VI., Chap. 3.

† Compare the title of Lexis' treatise, viz., *Zur Theorie der* MASSEN-*erscheinungen in der menschlichen Gesellschaft.*

and averages drawn from them are nearly stationary. By searching for the common properties contained in these numbers, we shall find the clue to this constancy. The following are among the numbers which do not undergo rapid change: birth, marriage, and death rates in districts of, say over a million inhabitants; death rates according to age or disease over larger areas; the numbers of the inhabitants of a great kingdom, even when subdivided by age and sex, the numbers of paupers, criminals, lunatics, afflicted; the consumption of certain commodities, the total income, the average wage, and total imports and exports (though here the constancy is not so apparent). These are all totals of many small items, the existence of each of which is determined independently and apparently by chance. Another class is to be found in meteorological measurements, such as annual rainfall, mean temperature, and mean barometric height, where the average or total is again drawn from the combination of many small independent variations or contributions. An allied class is found in such physical measurements as average height and weight.

The common property of great numbers.

It is not so easy to exemplify large numbers which are not constant. The total revenue which varies with each change in impost is an example. The number of a conscript army changes with the law controlling it, the number of volunteers with improved conditions of service, the area of the British Empire with each territorial extension, the volume of trade with a commercial inflation, the death rate with an epidemic. All these are changes where one cause has influenced many items at once in the same direction; but even here the underlying constancy arising from the multitudinous small independent causes is apparent.

This constancy, marvellous as it actually is, is generally accepted as a matter of course; and it is not the regularity but the occasional deflections which are the subject of comment. For instance, the death rate in London will hardly change, except regularly with the seasons, week by week through a series of years; and when an increase of 5 per 1,000 occurs in some week, the newspapers write of an influenza epidemic. The mean annual rainfall will for a long period be near its average; then a decrease of 5 inches excites remarks on a permanent change of climate. It is because this

Variation of great numbers.

regularity has become a matter of common experience that so little attention is generally given to it. A cursory inspection, however, of the records for a period of weeks or years of any of these numbers will show that the constancy is not absolute; that each rate varies through a great or small percentage, and, except that the variation seldom passes certain limits, without any apparent law. Thence at once rises the question, how are we to determine whether a given deviation is due to some general cause, such as an epidemic, a change of climate, or a new law, or is natural to the phenomena?

Illustration by the binomial expansion.

This question can only be answered by an appeal to the laws of probability. To take a numerical instance: suppose we are dealing with 1,000 men, each fifty years old, how many should we expect to die in the year? Fall back on former experience, and find what has been the average death rate under similar circumstances; this rate gives the number to be expected *à priori*, a great divergence appears from past experience to be improbable, and the greater the divergence the greater the improbability; an exact repetition of the average itself appears to be improbable; the question is, what divergence is to be expected? This is insoluble directly, but we can frame a hypothesis which throws light on the problem. Suppose the ascertained death rate to be 50 (per 1,000), and further suppose that the chance of death for each individual is $\frac{50}{1000}=\frac{1}{20}$. Then it is easily determined by the rules of algebraic probability that the successive terms in the expansion by the binomial theorem of $\left(\frac{19}{20}+\frac{1}{20}\right)^{1000}$ represent respectively the chances that exactly 0, 1, 2, 3 . . . of the persons die; $\left(\frac{19}{20}\right)^{1000}$ is the chance that none die, $\left(\frac{19}{20}\right)^{999}\left(\frac{1}{20}\right)$ is the chance that one assigned individual only dies, $1000\times\left(\frac{19}{20}\right)^{999}\left(\frac{1}{20}\right)$ that only one unassigned individual dies, and so on. The death of exactly 50 is more probable than any other number, 49 very nearly as probable, 51 next. It is very soon apparent when the successive terms are calculated, that any great divergence from 50 is very improbable.

This conception, that all the men start with the same chance of death, or, in a more developed form, that their chances of

death are grouped about an average $\frac{1}{20}$, satisfies the *à priori* conditions of the problem, and clearly leads to results which correspond roughly at any rate with experience; but the justice of the conception cannot be deduced *à priori*, for it is universally the case with any hypothesis as to probability, that conformity with experience is the only justification for the hypothesis. If it is true, we should find that when the records of many such generations of 1,000 men were examined, the divergences from the average were grouped in the way shown by the algebraic calculation. The records for this particular examination are not extant; but in the sequel some records will be given where experience marches with theory, and references will be given to books where others may be found; though it may be said at once that the agreement is not perfect, and that there are indications that the law is not so simple as that already suggested.

Process justified by its results.

Consider the supposition that the chance of a death within a year is $\frac{1}{20}$. When we say that the chance of an event is $\frac{1}{20}$, we mean that if the circumstances connected with it recurred again and again, the event would occur on an average once for each twenty such recurrences.* Thus if a die with six regular faces is thrown again and again, the different faces tend to come uppermost with equal frequency. As a matter of fact, each of the six would probably not be found once in each six throws, nor exactly two of each in twelve throws; but, in the long run, it is a matter of experience that the numbers of times each of the six faces come uppermost tend to be equal. Suppose an experiment, for the success of which the chance is $\frac{1}{20}$, to be performed again and again. In 200 attempts from 8 to 12 successes may be obtained; in 2,000 the proportion of successes to attempts will probably be nearer $\frac{1}{20}$, say 94 to 106 successes; in 2,000,000 yet nearer. Now suppose 1,000 experiments to be made; as we have seen, exactly 50 successes are not to be expected: but let 1,000 after 1,000 be tried; sometimes more, sometimes less than 50 successes will be obtained; and as the series continues the general average will tend nearer and nearer to 50.

The meaning of a numerical chance.

* *Logic of Chance*, 3rd Edition, pp. 4, 5.

Still postponing the examination of the exact grouping of such numbers about their average, let us examine further the nature of the argument. Suppose we are given a series of large numbers or rates, measuring similar quantities year after year; we shall find, when they are grouped according to their distance from their average, that the further from the average the fewer are the instances. In most cases we cannot work backwards to a number of individuals, each of whom has an equal chance of furnishing an event, but we can examine this grouping, notice how far the numbers are from their average, and so on; in many cases we shall find that these divergences conform to a definite law, the law of error, which is obeyed by all great numbers coming from series of experiments as just described. The point to notice specially here is, that correspondence to this regular law of divergence is natural, and it is for discrepancies that we need seek a reason. It is improbable, it is impossible, that great numbers should remain absolutely constant; from the nature of the case there must be variation; in very many cases the natural variation, the variation to be expected *à priori*, is that in accordance with the law of error. This is so with those great numbers which are the sum of very many items, in favour of the existence of each of which there is a definite chance, or, more generally, the existence of each of which may be influenced by many independent causes each of limited effect.

Relation to law of error.

A slight confusion may arise from the use of the words cause and chance in this statement; this can be removed by eliminating the word chance. We say a thing happens by chance, when its occurrence is influenced by many independent causes whose separate effects we cannot trace, as when we draw a card from a thoroughly shuffled pack. Now if we consider a man's death from the point of view of an insurance office, we regard the man as of normal health and constitution, and liable to all the latent diseases, the accidents, and the epidemics, from which experience shows men suffer; we cannot trace the incipient development of a disease, nor foretell the chain of events which lead to an accident. We then speak simply of the chance of death within a certain period, and say experience shows it to be (*e.g.*) $\frac{1}{20}$, and, regarding the peculiarities of a particular man as unknown, we say that his chance of death is

Cause and chance.

$\frac{1}{20}$. Generalizing, any group of men, each of the given age and in the given circumstances, is composed of individuals for each of whom the chance of death is $\frac{1}{20}$. Now, go behind the idea of chance to that of cause. Each death is the result of some particular event, or, to speak more correctly, is due to the action of a complex of many causes; all these untraceable causes produce on an average one death among 20 living; the statement of the numerical chance is merely the summary of these effects. To say, then, that the number of deaths to be expected among 1,000 is the same as the number of successes to be expected in 1,000 attempts, the chance of success in each of which is $\frac{1}{20}$, is not inconsistent with saying that the number of deaths is determined by the action of a multitude of causes none of which by itself produces a great effect. In either case the laws of great numbers will be found to apply. The use of the intermediate numerical chance only facilitates calculation.

Now suppose that a new cause is suddenly introduced, or the action of one of the causes is intensified (say, by an epidemic), and at once the whole scheme of calculation is thrown out, and we get a result which does not correspond to the probability calculation; it is this non-correspondence which indicates the existence of a disturbing cause.

Effect of a predominant cause.

Since the distribution in accordance with the curve of error is the result which may be expected *à priori*, whenever we are dealing with numbers generated in this way, it is clearly necessary to study this distribution before we can base any arguments on the variation of great numbers. When we have established the result which the independent action of a very great number of individually unimportant causes can produce, then, and not till then, we are in a position to consider the effect of a predominant cause. We may even be able to deduce the existence of such a cause, for if we find by examination that a divergence of more than 3 per cent. from an average is improbable, and in a particular case we have a divergence of 30 per cent., we are either in the presence of a very improbable event, or some external predominant cause has influenced our numbers.

SECTION II.—THE EQUATION OF THE CURVE OF ERROR.

The section can be omitted, but not without loss.

In this section it is proposed to develop the algebraic equation and properties of the curve of error, bringing them into relation with the other sections. It will be not impossible for non-mathematical readers to follow the great part of the argument of this branch of statistics without working through the mathematical proofs of the formulæ; and the book is so arranged that this section can be omitted. Other readers may turn this chapter through looking at the large type only, and notice the main lines of the argument. For any thorough student of statistics, however, the mathematical proofs, which are so simplified in this chapter as not to involve the integral calculus at all, and the differential calculus only for two small points, are essential. In this section an acquaintance with algebra up to and including the exponential and logarithmic series is assumed. Starting from that point, the main formulæ relating to the curve of error are deduced.

ELEMENTARY THEOREMS IN PROBABILITY.

DEFINITION.—If an event can happen in m ways, and fail in $\overline{n-m}$ ways, and all these ways are equally likely to occur, then $\frac{m}{n}$ is the probability of its occurrence.

Let $\frac{m}{n} = p$, and $\frac{n-m}{n} = q$: then $p + q = 1$. q is the chance that event will not occur. The odds in favour are p to q, those against are q to p.

E.g., the chance that a card, drawn at random from a full pack, is a spade $= \frac{13}{52} = \frac{1}{4}$.

Theorem.—If $p_1 p_2$ are the chances of two independent events, then $p_1 \times p_2$ is the chance that both will occur.

Suppose that $p_1 = \frac{m_1}{n_1}$, $p_2 = \frac{m_2}{n_2}$.

The first event may be expected m_1 times in n_1 trials, or $m_1 n_2$ times in $n_1 n_2$ trials.

The second event may be expected m_2 times in n_2 trials, or m_1m_2 times in m_1n_2 trials.

Hence the second event will occur at same time as the first m_1m_2 times in n_1n_2 trials; that is, the chance of the double event is $\frac{m_1m_2}{n_1n_2} = p_1p_2$.

EXAMPLES.—*Independent events.*—The chance that two sixes will be thrown with a pair of dice $= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

Dependent events.—The chance that three cards taken in succession from the same pack shall prove to be ace, king, and queen of the same suit in any order is $\frac{12}{52} \times \frac{2}{51} \times \frac{1}{50} = \frac{1}{5525}$; for the chance that the first card drawn is an ace, king, or queen is $\frac{12}{52}$; supposing it to be queen, the chance that ace or king of same suit follows is $\frac{2}{51}$; and the chance that the third draw gives the remaining card is $\frac{1}{50}$.

The chance that 13 cards taken at random from a complete pack will contain 8 spades and 5 clubs is $\frac{{}^{13}C_8 \times {}^{13}C_5}{{}^{52}C_{13}}$;* for 8 spades can be chosen in ${}^{13}C_8$ ways, 5 clubs in ${}^{13}C_5$ ways, and the hand may contain any such group of spades with any such group of clubs; hence the numerator given corresponds to m of the definition given above; also there are ${}^{52}C_{13}$ equally likely possible hands of 13 cards, so that the denominator given corresponds to n.

Theorem.—If n coins are placed at random on a table, the chance that r will show heads and the rest tails is $\frac{{}^nC_r}{2^n}$.

For suppose there are n places to be filled each with a coin:—

The first may show head or tail, two ways.

The second may show head or tail, two ways.

The first two places may therefore be filled in 2×2 ways.

The n places may similarly be filled in 2^n ways.

Now r of these places can be chosen in ${}^nC_r$ ways; and to each such selection corresponds one arrangement in which these r places are filled

---

* ${}^nC_r$ or ${}_nC_r$ is written for $\frac{n!}{n-r!\,r!}$, the numbers of combinations of n things r at a time; and ${}^nP_r$ or ${}^nP_r$ is written for $\frac{n!}{n-r!}$, the number of Permutations of n things r at time.

with heads and the rest with tails (and many other arrangements not giving this result).

Hence out of 2^n possible arrangements, ${}^nC_r$ give the result.

Aliter.—Consider the product $(h_1 + t_1)\ (h_2 + t_2)\ \ldots\ (h_n + t_n)$. Any term of this product, *e.g.*, $h_1\ t_2\ t_3\ h_4\ t_5\ \ldots\ t_{n-1}\ h_n$ corresponds to one arrangement of the n coins.

The number of arrangements containing r heads and $\overline{n-r}$ tails is the same as the number of terms containing r h's and $\overline{n-r}$ t's.

This number is the same as the coefficient of $h^r t^{n-r}$ is the expansion by the binomial theorem of $(h+t)^n$, which is obtained from the product above by writing h for h_1, h_2, &c., and t for t_1, t_2, &c.

$$(h+t)^n = h^n + {}^nC_1.\ h^{n-1}\ t + \ldots + {}^nC_r\ h^{n-r}\ t^r + \ldots + t^n.$$

Hence the number of arrangements producing the required result is ${}^nC_r$. The total number of possible arrangements is the sum of the coefficients in this expansion; this is found by putting $h = t = 1$.

$$(1+1)^n = 1 + {}^nC_1 + \ldots + {}^nC_r + \ldots + 1.$$

Hence total number $= 2^n$.

Notice that ${}^nC_r = {}^nC_{n-r}$.

EXAMPLE.—The coefficients in the expansion of $(1+1)^{32}$ are as follows:—

${}^{32}C_0 = {}^{32}C_{32}$	=	1.	Corresponding chance * .0000000002
${}^{32}C_1 = {}^{32}C_{31}$	=	32.	,, .0000000074
${}^{32}C_2 = {}^{32}C_{30}$	=	496.	,, .0000001155
${}^{32}C_3 = {}^{32}C_{29}$	=	4,960.	,, .0000011555
${}^{32}C_4 = {}^{32}C_{28}$	=	35,960.	,, .0000083750
${}^{32}C_5 = {}^{32}C_{27}$	=	201,376.	,, .0000468800
${}^{32}C_6 = {}^{32}C_{26}$	=	906,192.	,, .0002110000
${}^{32}C_7 = {}^{32}C_{25}$	=	3,365,856.	,, .0007837000
${}^{32}C_8 = {}^{32}C_{24}$	=	10,518,300.	,, .0024490000
${}^{32}C_9 = {}^{32}C_{23}$	=	28,048,800.	,, .0065300000
${}^{32}C_{10} = {}^{32}C_{22}$	=	64,512,240.	,, .0150200000
${}^{32}C_{11} = {}^{32}C_{21}$	=	129,024,480.	,, .0300400000
${}^{32}C_{12} = {}^{32}C_{20}$	=	225,792,840.	,, .0525700000
${}^{32}C_{13} = {}^{32}C_{19}$	=	347,373,600.	,, .0808800000
${}^{32}C_{14} = {}^{32}C_{18}$	=	471,435,600.	,, .1097000000
${}^{32}C_{15} = {}^{32}C_{17}$	=	565,722,720.	,, .1317000000
${}^{32}C_{16}$	=	601,080,390.	,, .1400000000
		$2^{32} = 4,294,967,296.$	

The table just given shows that when 32 coins are placed on a table at random, the chance that 16 heads and 16 tails shall appear

* Obtained by dividing each term by 2^{32}.

is .14, while it is more likely that either 15 heads (and 17 tails) or 15 tails (and 17 heads) will be found, the united chances for these being .2634. The chance that the divergence from equal division shall not exceed 2 (*i.e.*, that there shall be at least 14 of each) is $.1400 + 2 \times (.1317 + .1097) = .6228$; the chance that there shall be as many as 27 of one kind is only .00011, *i.e.*, 1 in 9,000.

THE BINOMIAL EXPANSION.

The following table from Quetelet's *Lettres sur la Theorie des Probabilités*, p. 375, shows a similar calculation when the index is 999 instead of 32. For instance, ${}^{999}C_{500}\ (\frac{1}{2})^{999} = .025225$, the first quantity in Column 3.

As the index of the binomial expansion is continually increased, the grouping of the figures takes a definite shape. The curve so obtained when the index is indefinitely great is called the *curve of error*.

In the diagram at the end of the book, the line $A_2\ F_1\ F_2$ represents the first half of the coefficients of $(a+b)^4$; the line $A_2\ G_1\ G_2\ G_3\ G_4\ G_5$ represents the coefficients of $(a+b)^{10}$; the line $A_2\ H_1\ H_2\ H_3 \ldots H_9$ represents the coefficients of $(a+b)^{32}$, and $A_0\ A_1\ A_2\ A_3\ A_4$ is the curve of error. To fit these jagged lines to the curve of error, the maximum coefficient is represented in each case by the line $O\ A_2$, and ordinates are drawn at equal intervals proportional to the other coefficients; the tops of these ordinates are then joined by straight lines. The interval between successive ordinates is decided by the consideration that the area included between any chosen ordinate, say $H_2\ P_1$, the base $P_1\ O$, the maximum ordinate $O\ A_2$, and the line $H_2\ H_1\ A_2$ shall be the same fraction of the whole area $A_2\ O \ldots H_9 \ldots A_2$, as the part of the area of the limiting curve of error cut off by the same ordinate is of the whole area bounded by $O\ A_2$, $O\ X$ and the curve.

The algebraic determination of this limit is given on pp. 275 *seq.*

Suppose now that one ball is taken out of each of n bags, each containing m_1 white and m_2 black balls, the chance that r will be white and $n-r$ black is—

$$ {}^nC_r\, p^r.\ q^{n-r}, \text{ where } p = \frac{m_1}{m} \text{ and } q = \frac{m_2}{m} \text{ and } m = m_1 + m_2. $$

For r bags may be selected in ${}^nC_r$ ways. The chance that each of

SCALE OF PRECISION.

999 balls are drawn from a bag containing equally great numbers of black and white balls.

Column 1 gives number of each colour.
" 2 gives rank of deviation from equality.
" 3 gives probability that balls will be drawn in proportion given in Column 2.
" 4 gives probability that deviation from equality will not be greater than that of given rank.

1. Groups of		2.	3. Scale of Probability.	4. Scale of Precision.	1. Groups of		2.	3. Scale of Probability.	4. Scale of Precision.
White.	Black.	Rank.	Probability that such a group will be drawn.	Sum of probabilities starting from most probable.	White.	Black.	Rank.	Probability that such a group will be drawn.	Sum of probabilities starting from most probable.
499	500	1	.025225	.025225	459	540	41	.0009458	.495278
498	501	2	.025124	.050349	458	541	42	.0008024	.496081
497	502	3	.024924	.075273	457	542	43	.0006781	.496759
496	503	4	.024627	.099900	456	543	44	.0005707	.497329
495	504	5	.024236	.124136	455	544	45	.0004784	.497808
494	505	6	.023756	.147892	454	545	46	.0003994	.498207
493	506	7	.023193	.171085	453	546	47	.0003321	.498539
492	507	8	.022552	.193637	452	547	48	.0002750	.498814
491	508	9	.021842	.215479	451	548	49	.0002268	.499041
490	509	10	.021069	.236548	450	549	50	.0001863	.499227
489	510	11	.020243	.256791	449	550	51	.0001525	.499380
488	511	12	.019372	.276163	448	551	52	.0001242	.499504
487	512	13	.018464	.294627	447	552	53	.0001008	.499605
486	513	14	.017528	.312155	446	553	54	.0000815	.499686
485	514	15	.016573	.338728	445	554	55	.0000656	.499752
484	515	16	.015608	.344335	444	555	56	.0000526	.499804
483	516	17	.014640	.358975	443	556	57	.0000421	.499847
482	517	18	.013677	.372652	442	557	58	.0000334	.499880
481	518	19	.012726	.385378	441	558	59	.0000265	.499906
480	519	20	.011794	.397172	440	559	60	.0000209	.499927
479	520	21	.010887	.408060	439	560	61	.0000164	.499944
478	521	22	.010008	.418070	438	561	62	.0000128	.499957
477	522	23	.009166	.427236	437	562	63	.0000100	.499967
476	523	24	.008360	.435595	436	563	64	.0000077	.499974
475	524	25	.007594	.443189	435	564	65	.0000060	.499980
474	525	26	.006871	.450060	434	565	66	.0000046	.499985
473	526	27	.006191	.456251	433	566	67	.0000035	.499988
472	527	28	.005557	.461809	432	567	68	.0000027	.4999912
471	528	29	.004968	.466776	431	568	69	.0000021	.4999933
470	529	30	.004423	.471199	430	569	70	.0000016	.4999948
469	530	31	.003922	.475122	429	570	71	.0000012	.4999960
468	531	32	.003464	.478586	428	571	72	.0000009	.4999969
467	532	33	.003047	.481633	427	572	73	.0000007	.4999976
466	533	34	.002670	.484304	426	573	74	.0000005	.4999981
465	534	35	.002330	.486634	425	574	75	.0000004	.4999984
464	535	36	.002025	.488659	424	575	76	.0000003	.4999987
463	536	37	.001753	.490412	423	576	77	.0000002	.4999989
462	537	38	.001512	.491924	422	577	78	.00000014	.4999990
461	538	39	.001298	.493222	421	578	79	.00000011	.4999991
460	539	40	.001110	.494332	420	579	80	.00000004	.4999992

By means of this scale the binomial 999, practically equivalent to curve of error, can be fitted to and compared with any series of observations.

these will yield a white ball is $p \times p \times p \times \ldots$ to r factors, *i.e.*, p^r; the chance that each of the other $n-r$ bags will yield a black ball is q^{n-r}; hence required chance is as stated.

Aliter.

Call the white balls in the first bag ${}_1w_1, {}_1w_2, \ldots {}_1w_{m_1}$;
" black " " ${}_1b_1, {}_1b_2, \ldots {}_1b_{m_2}$;
" white balls in the second bag ${}_2w_1, {}_2w_2, \ldots {}_2w_{m_1}$;
" black " " ${}_2b_1, {}_2b_2, \ldots {}_2b_{m_2}$;

and so on; then all possible arrangements are represented by the individual terms of the product—

$$({}_1w_1 + {}_1w_2 + \ldots + {}_1w_{m_1} + {}_1b_1 + {}_1b_2 + \ldots + {}_1b_{m_2}) \times$$
$$({}_2w_1 + {}_2w_2 + + {}_2w_{m_1} + {}_2b_1 + {}_2b_2 + \ldots + {}_2b_{m_2}) \times \times n \text{ factors;}$$

e.g., the term ${}_1w_1 \cdot {}_2b_1 \cdot {}_3b_5 \cdot {}_4w_7 \ldots {}_nb_{m_2}$ represents one group. A w will occur r times and a b the remaining $n-r$ times as often as the term $w^r b^{n-r}$ occurs in the binomial expansion of $(m_1 w + m_2 b)^n$ [where all the w's are put as w, and all the b's as b]. The coefficient of $w^r b^{n-r}$ in this expansion is ${}^nC_r \cdot m_1^r \cdot m_2^{n-r}$. Total number of possible arrangements is m^n. Hence required chance is—

$$\frac{{}^nC_r \cdot m_1^r \cdot m_2^{n-r}}{m^n} = {}^nC_r \cdot \left(\frac{m_1}{m}\right)^r \left(\frac{m_2}{m}\right)^{n-r} = {}^nC_r \cdot p^r \cdot q^{n-r}.$$

E.g., to find the probable number of sixes in n throws of dice. Here $m=6, m_1=1, m_2=5, p=\frac{1}{6}, q=\frac{5}{6}$.

Probability of r sixes $= {}^nC_r \left(\frac{1}{6}\right)^r \left(\frac{5}{6}\right)^{n-r}$

Suppose $n = 12$. Total number of possible arrangements is $6^{12} = 2,176,782,336$.

12 sixes occur		${}_{12}C_{12} \cdot 1^{12} \cdot 5^0$ times =	1
11	"	${}_{12}C_{11} \cdot 1^{11} \cdot 5^1$ "	60
10	"	${}_{12}C_{10} \cdot 1^{10} \cdot 5^2$ "	1,650
9	"	${}_{12}C_9 \cdot 1^9 \cdot 5^3$ "	27,500
8	"	${}_{12}C_8 \cdot 1^8 \cdot 5^4$ "	309,375
7	"	${}_{12}C_7 \cdot 1^7 \cdot 5^5$ "	2,475,000
6	"	${}_{12}C_6 \cdot 1^6 \cdot 5^6$ "	14,437,500
5	"	${}_{12}C_5 \cdot 1^5 \cdot 5^7$ "	61,875,000
4	"	${}_{12}C_4 \cdot 1^4 \cdot 5^8$ "	193,359,375
3	"	${}_{12}C_3 \cdot 1^3 \cdot 5^9$ "	429,687,500
2	"	${}_{12}C_2 \cdot 1^2 \cdot 5^{10}$ "	644,531,250
1	"	${}_{12}C_1 \cdot 1^1 \cdot 5^{11}$ "	585,937,500
0	"	${}_{12}C_0 \cdot 1^0 \cdot 5^{12}$ "	244,140,625
		Total	2,176,782,336

The most probable number of sixes is 2, of which the chance is about $\frac{2}{7}$. In four-fifths of the trials there will probably be 1, 2, or 3 sixes.

E.g., to investigate whether drunkenness occurs chiefly on night of pay-day (suppose Saturday). If maximum number of convictions in a week is on Saturday in 10 weeks out of 12 selected at random, we have an event whose probability is only $\frac{1650}{2176782336} = \frac{1}{1300000}$ (about), if the position of the day in the week had nothing to do with it.

It must be noticed that the probability that event will occur 10 times out of 12 on any the same week-day is much greater, viz., $\frac{1}{220000}$.

Probability that event will occur *at least* 10 times is $\frac{1+60+1650}{6^{12}}$ $= \frac{1711}{2176782336}$ which is much the same as before.

Similarly any questions depending on the occurrence of an event in the same month may be worked out. In this case $m_1 = 1$, $m_2 = 11$, $n =$ number of years investigated.

If a bag contains m balls of different colours, m_1 white, m_2 red, m_3 green, &c., the probability that r_1 white, r_2 red, r_3 green, &c., will occur is coefficient of $p_1^{r_1} p_2^{r_2} p_3^{r_3} \ldots$ in expansion of $(p_1+p_2+p_3+)^n$ by the multinomial theorem, where $p_1 = \frac{m_1}{m}$, $p_2 = \frac{m_2}{m}$, &c.

Notice that the probability of an event, *if* it was a chance occurrence, is not the same as the probability that the event *was* a chance occurrence.

If 13 trumps appeared in the same hand, we could not say that the chances were (${}^{51}C_{12} =$) 158,753,389,900 to 1 that the hand was "faked," but we should have strong though incommensurable evidence on the point.

Deduction of Equation of Curve of Error.

We can now proceed to the determination of the equation of the curve of error.

The chance of r successes is greatest when r is the greatest integer in pn; this is found by the ordinary method of determining the maximum term in a binomial expansion.

Let P be this maximum value $= {}^nC_{pn} \cdot p^{pn} q^{qn}$, making the supposition for brevity that pn is integral, which will not affect the proof.

$$= \frac{\lfloor n}{\lfloor pn \lfloor qn} p^{pn} q^{qn}, \text{ for } pn + qn = n.$$

Let P_x be chance of $pn + x$ white balls.

Then
$$P_x = P \times \left(\frac{p}{q}\right)^x \times \frac{qn.\,(qn-1)\ldots(qn-x+1)}{(pn+1)\,(pn+2)\ldots(pn+x)}$$
$$= P \times \frac{1.\left(1-\frac{1}{qn}\right)\left(1-\frac{2}{qn}\right)\ldots\left(1-\frac{x-1}{qn}\right)}{\left(1+\frac{1}{pn}\right).\left(1+\frac{2}{pn}\right)\ldots\left(1+\frac{x}{pn}\right)}$$

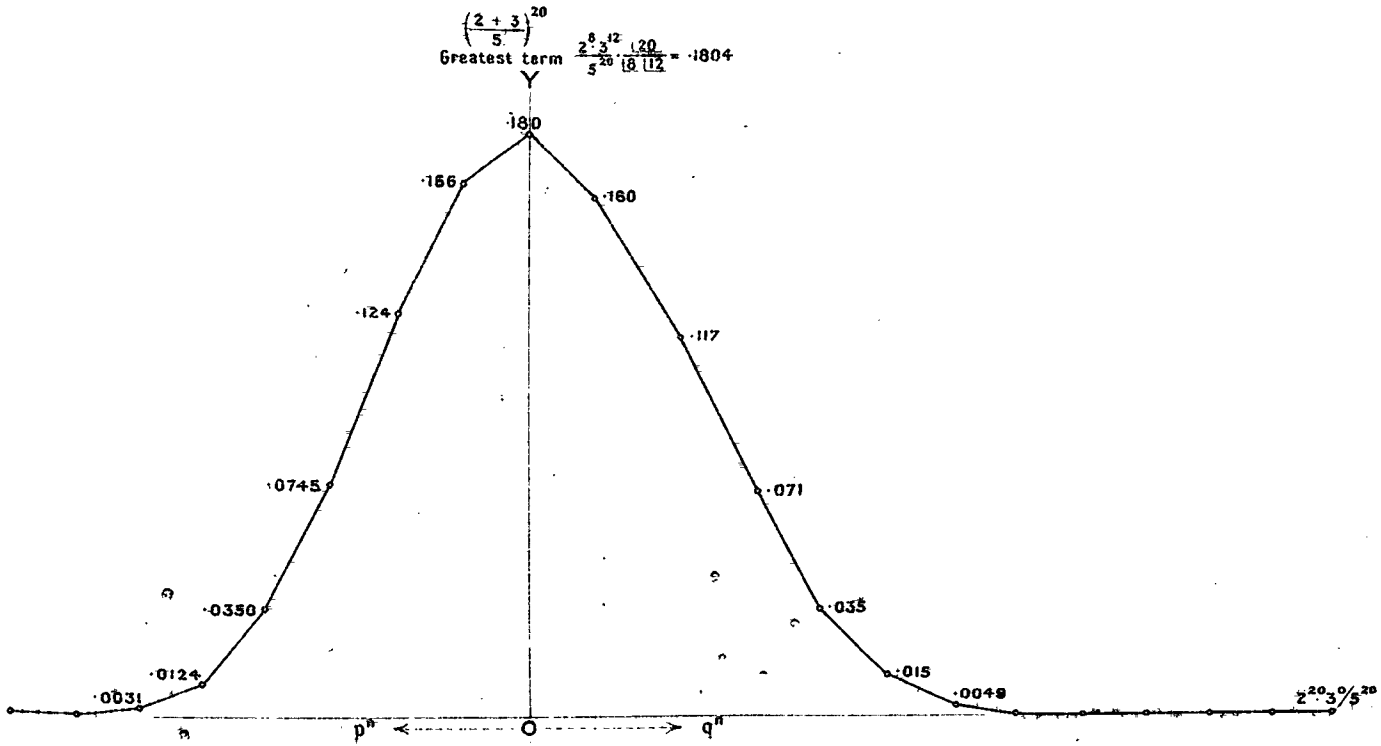
Taking logarithms of both sides—

$$\log P_x = \log P + \log\left(1-\frac{1}{qn}\right) + \log\left(1-\frac{2}{qn}\right) + \ldots + \log\left(1-\frac{x-1}{qn}\right)$$
$$- \log\left(1+\frac{1}{pn}\right) - \log\left(1+\frac{2}{pn}\right) - \ldots - \log\left(1+\frac{x-1}{pn}\right) - \log\left(1+\frac{x}{pn}\right)$$
$$= \log P - \left(\frac{1}{qn} + \frac{1}{2}\cdot\frac{1}{(qn)^2} + \right) - \left(\frac{2}{qn} + \frac{1}{2}\cdot\left(\frac{2}{qn}\right)^2 + \right) - \ldots$$
$$- \left(\frac{x-1}{qn} + \frac{1}{2}\left(\frac{x-1}{qn}\right)^2 + \right)$$
$$- \left(\frac{1}{pn} - \frac{1}{2}\cdot\frac{1}{(pn)^2} + \right) - \left(\frac{2}{pn} - \frac{1}{2}\left(\frac{2}{pn}\right)^2 + \right) - \ldots - \left(\frac{x}{pn} - \frac{1}{2}\left(\frac{x}{pn}\right)^2 + \right)$$
$$= \log P - \frac{1+2+\ldots+\overline{x-1}}{qn} - \frac{1^2+2^2+\ldots+\overline{x-1}^2}{2q^2n^2} - \ldots\ldots$$
$$- \frac{1+2+\ldots+x}{pn} + \frac{1+2^2+\ldots+x^2}{2p^2n^2} - + \ldots$$
$$= \log P - \frac{x(x-1)}{2qn} - \frac{x(x+1)}{2pn} - \frac{(x-1).x.(2x-1)}{12q^2n^2} + \frac{x(x+1)(2x+1)}{12p^2n^2} - +$$

Now when n, the number of balls in each bag, is very great, pn and qn are also very great if neither p nor q are very small; x, the divergence from pn, ranges from O to qn on the positive side of the maximum, and O to $-pn$ on the negative side. The chance of so great a divergence as $-pn$ or qn is very small. The chance of a small divergence, such as $x = 1, 2, 3 \ldots$ is very nearly equal to the maximum chance P. For instance, if $x = 3$—

$$P_3 = P \times \left(\frac{p}{q}\right)^3 \times \frac{qn\,(qn-1)\,(qn-2)}{(pn+1)\,(pn+2)\,(pn+3)}$$
$$= P\left(1-\frac{1}{nq}\right)\left(1-\frac{2}{nq}\right)\left(1+\frac{1}{pn}\right)^{-1}\left(1+\frac{2}{pn}\right)^{-1}\left(1+\frac{3}{pn}\right)^{-1}$$
$$= \text{(expanding each term) } P \times \left[1 - \frac{3}{nq} - \frac{6}{np} + \text{terms involving } \frac{1}{n^2}\right].$$

SYMMETRY OF THE BINOMICAL CURVE NEAR ITS GREATEST ORDINATE.



To face page 217.

Hence, in order that P_x may have at once a finite value and one with a finite divergence from P, x must be very great compared with unity, but small compared with n.

Re-write the above equation for P_x, neglecting $\left(\frac{x}{n}\right)$,

$$\log P_x = \log P - \frac{x^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right) + \frac{x^3}{6n^2}\left(\frac{1}{p^2}-\frac{1}{q^2}\right) - \frac{x^4}{n^3}\left(\quad\right) \text{ &c.}$$

If $\frac{x^2}{n}$ were negligible, $\log P_x = \log P$.

If $\frac{x}{n}$ were finite, and therefore x infinite, both $\frac{x^3}{n^2}$ and $\frac{x^4}{n^3}$ would be infinite.

That part of the resulting curve, which shows finite curvature, is found by assuming that x and n are infinite, but $\frac{x^2}{n}$ finite.

[The general argument is similar to that used for constructing the finite part of a parabola on a finite scale, for there $\frac{y^2}{x}$ is finite.]

On this hypothesis $\frac{x^3}{n^2} = \left(\frac{x^2}{n}\right)^2 \cdot \frac{1}{x}$, $\frac{x^4}{n^3} = \left(\frac{x^2}{n}\right)^2 \cdot \frac{1}{n}$, and these and presumably further terms are infinitesimal. The equation of the finite part of the curve is therefore—

$$\log P_x = \log P - \frac{x^2(p+q)}{2npq}$$

$$\text{or} \qquad P_x = P.e^{-\frac{x^2}{2npq}}, \text{ since } p+q = 1.$$

Writing y for P_x, $y = Pe^{-\frac{x^2}{2npq}}$.

The curve is horizontal near the maximum ordinate, for $P_x = P$ when x is small, and extends to infinity in both directions, the axis of x being an asymptote, for when $x = \pm\infty$, y is zero.

When $\frac{x^3}{n^2}$ is negligible, the curve is very approximately symmetrical; this symmetry may be shown to extend over the finite part of the curve, when n is large.* The annexed diagram illustrates the extent of the asymmetry for a small value of n.

* By considering the values of the various quantities in relation to the table on p. 281.

Relation of Curve of Error to Statistics.

The following example shows how Quetelet fitted his figures to given observations: *—

Chest Measurements of Scotch Soldiers.

1. Chest Measurement, Inches.	2. No. of Men.	3. Proportional Nos.	4. No. between given Measurement and Mean.	5. Rank in Scale of Precision.	6. Calculated Rank of Measurement.	7. Precision of Calculated Rank.	8. Calculated No. of Observations to each Measurement.	9. Differences between Columns 3 and 8.
33	3	5	5,000	...	...	5,000	7	2
34	18	31	4,995	+52	50	4,993	29	2
35	81	141	4,964	42.5	42.5	4,964	110	31
36	185	322	4,823	33.5	34.5	4,854	323	1
37	420	732	4,501	26.0	26.5	4,531	732	0
38	749	1,305	3,769	18.0	18.5	3,799	1,333	28
39	1,075	1,867	2,464	10.5	10.5	2,466	1,838	29
	...	...	597	+2.5	2.5	628	...	...
40	1,079	1,882	1,285	−5.5	5.5	1,359	1,987	105
41	934	1,628	2,913	13	13.5	3,034	1,675	47
42	658	1,148	4,061	21	21.5	4,130	1,096	52
43	370	645	4,706	30	29.5	4,690	560	85
44	92	160	4,866	35	37.5	4,911	221	61
45	50	87	4,953	41	45.5	4,980	69	18
46	21	38	4,991	49.5	53.5	4,996	16	22
47	4	7	4,998	−56	61.8	4,999	3	4
48	1	2	5,000	...	...	5,000	1	1
...	5,738	10,000	...	...	...	...	10,000	488

The chest measurements of 5,738 soldiers were ranged in order of magnitude, and the numbers of men at each measurement placed in column 2 against the corresponding number of inches in column 1. Column 3 gives numbers proportional to those in column, such that their sum is 10,000. It is assumed that there are 5,000 cases on each side of the (unknown) mean; then 5,000 cases occur between 33 inches and the mean, 4,995 between 34 inches and mean, and so on, till we find (in column 4) 597 cases between 39 inches and mean. Similarly 1,285 occur between 40 inches and the mean.

Referring now to the scale of precision, we find that 4,995 cases corresponds to rank 52; 4,964 to rank 42.5, and so on. The numbers of these ranks are placed in column 5.

Now if the observations fitted the curve exactly the distances

* *Ibid.*, p. 400.

between two ranks corresponding to two successive inches should be always the same. This is not exactly the case: 34 inches corresponds to rank 52; 35 inches to 9.5 ranks lower, 42.5; 36 to 9 ranks lower, and so on. It is necessary to assume some regular interval, which will show as close a correspondence as possible between theory and observation. It is assumed that a difference of 1 inch corresponds to 8 ranks, and column 5 is "smoothed" into column 6 on this hypothesis. The process is then reversed; against each rank in column 6 is placed in column 7 the corresponding number from the scale of precision. Column 8 is then calculated from column 7 in the reverse way to that by which column 4 is reckoned from column 3. The closeness of the resemblance between columns 8 and 3 shows how nearly the measurements fit the theory. In column 9 are placed the differences between the numbers in columns 3 and 8; the percentage which the sum of these differences is of the total number (10,000), is a measure of the fit. If the observations are plotted out in the same way as the later figures which form the L_1, L_2 . . . figure on the diagram at the end of the book, this ratio of the sum of the differences to the whole number, is a rough measure of the ratio of the sum of the areas included between the lines L_1 L_2 . . . and the curve of error $A_1A_2A_3$. In the case just discussed the misfit is 4.88 per cent.

The following seems a better method of estimating the fit, for it is less dependent on the accidental divergences caused by the particular interval of measurement taken. Construct a figure to represent the scale of precision on p. 273; fit as closely as possible to this another figure, whose ordinates represent the numbers "at or above" given measurements represented by the abscissæ; the whole curve will be nearly the shape of the figure facing p. 155, when the smoothed curve may stand for the scale of precision, symmetrical about its median, and the original jagged line for the observations. The closeness of the jagged to the smoothed line shows the fit; and this will not be altered by a slight shifting of the inch or shilling limits we adopt, which in the other method often makes a great difference to the regularity * in discontinuous observations. Moreover, it is not

* *E.g.*, two great numbers at 29s. 9d. and 30s. 3d. respectively will both be in the same group if our limits are "29s. 6d. and not so much as 30s. 6d.," &c., but in different groups if the limits are "29s. and not so much as 30s.," &c.

necessary by the latter method to sort the observations into groups at all.

We cannot deduce this equation from the most general hypotheses (stated on p. 303, *infra*) without the use of the integral calculus. It is, however, more convenient to use a scale of precision evaluated from the equation of the curve as obtained in other ways. In the next table the numbers under x correspond to Quetelet's "ranks," but the divergence taken as unity corresponds to the quantity $\sqrt{2pqn}$ which is rank 22.35 . . . (for $\sqrt{2\times\frac{1}{2}\times\frac{1}{2}\times 999}=22.35\ldots$) in Quetelet's scale. The quantities under $F(x)$ correspond to those in the earlier scale of precision; so that against any value of x is found the chance that an observation shall be between the average and x. The figures are adapted from Lexis' *Massenerscheinungen*, pp. 93, 94, and Quetelet, *ibid.*, p. 389. This table and Quetelet's can be used indifferently; they yield very nearly the same results.

The diagram at the end of the book shows the relation of the binomial with index, 4, 10, and 32, to the curve of error.

Note.—For the evaluation of large factorials met with in work of this sort, Stirling's approximate formula is useful. This is—

$$(m-1)! = m^{m-\frac{1}{2}}.(2\pi)^{\frac{1}{2}}.e^{-m+\frac{1}{12m}-\frac{1}{360m^3}}$$

The following adaptation will serve for four significant figures:—

$$\log_{10} n! = (n+\tfrac{1}{2})\log_{10}(n+1) - .0352 + \left(-n+\frac{1}{12(n+1)}\right)\times .43439$$

VALUES OF F (x) FOR DIFFERENT VALUES OF x, WHERE

$$F(x) = \frac{1}{\sqrt{\pi}} \int_0^x e^{-x^2} . dx.$$

x.	F(x).	x.	F(x).	x.	F(x).	x.	F(x).	x.	F(x).
.00	.000	.36	.195	.71	.342	1.06	.433	1.41	.477
.01	.006	.37	.200	.72	.346	1.07	.435	1.42	.478
.02	.011	.38	.205	.73	.349	1.08	.437	1.43	.478
.03	.017	.39	.209	.74	.352	1.09	.438	1.44	.479
.04	.023	.40	.214	.75	.356	1.10	.440	1.45	.480
.05	.028	.41	.219	.76	.359	1.11	.442	1.46	.481
.06	.034	.42	.224	.77	.362	1.12	.443	1.47	.481
.07	.039	.43	.228	.78	.365	1.13	.445	1.48	.482
.08	.045	.44	.233	.79	.368	1.14	.447	1.49	.482
.09	.051	.45	.238	.80	.371	1.15	.448	1.50	.483
.10	.056	.46	.242	.81	.374	1.16	.450	1.52	.484
.11	.062	.47	.247	.82	.377	1.17	.451	1.54	.485
.12	.067	.48	.251	.83	.380	1.18	.452	1.56	.486
.13	.073	.49	.256	.84	.383	1.19	.454	1.58	.487
.14	.078	.50	.260	.85	.385	1.20	.455	1.60	.488
.15	.084	.51	.265	.86	.388	1.21	.456	1.62	.489
.16	.090	.52	.269	.87	.391	1.22	.458	1.64	.490
.17	.095	.53	.273	.88	.393	1.23	.459	1.66	.491
.18	.100	.54	.277	.89	.396	1.24	.460	1.68	.491
.19	.106	.55	.282	.90	.398	1.25	.461	1.70	.492
.20	.111	.56	.286	.91	.401	1.26	.463	1.72	.493
.21	.117	.57	.290	.92	.403	1.27	.464	1.74	.493
.22	.122	.58	.294	.93	.406	1.28	.465	1.76	.494
.23	.128	.59	.298	.94	.408	1.29	.466	1.78	.494
.24	.133	.60	.302	.95	.410	1.30	.467	1.80	.495
.25	.138	.61	.306	.96	.413	1.31	.468	1.82	.495
.26	.143	.62	.310	.97	.415	1.32	.469	1.84	.495
.27	.149	.63	.314	.98	.417	1.33	.470	1.86	.496
.28	.154	.64	.317	.99	.419	1.34	.471	1.88	.496
.29	.159	.65	.321	1.00	.421	1.35	.472	1.90	.496
.30	.164	.66	.325	1.01	.423	1.36	.473	1.92	.496
.31	.169	.67	.328	1.02	.425	1.37	.474	1.94	.497
.32	.175	.68	.332	1.03	.427	1.38	.475	1.96	.497
.33	.180	.69	.335	1.04	.429	1.39	.475	1.98	.497
.34	.185	.70	.339	1.05	.431	1.40	.476	2.00	.498
.35	.190							2.05	.498

x.	F(x).
2.20	.49907
2.50	.49980
3.00	.499,989
4.00	.499,999,992
5.00	.499,999,999,999,2 . .

SPECIAL POINTS ON THE CURVE.

Before we can show how to fit observations to this table, we must consider the equation of the law of error more closely.

DEFINITION.—The *probable error* of a series of observations

is that divergence from their mean on either side, within which exactly half the observations lie. This quantity is more appropriately called the *Quartile Deviation.**

In the scale of precision the corresponding number is .25, either above or below the mean; the corresponding rank in Quetelet's scale is about 10.7. In the table just given, the value of $F(x)$ which corresponds to the probable error is .25, and the corresponding value of x is calculated to be .47694, a quantity usually designated by ρ.

An approximation to the probable error for a given series of observations is obtained by arranging all the observations in order of magnitude; marking the magnitude, say α, above which 25 per cent. of the observations lie, and the magnitude, say β, below which 25 per cent. lie. Half the difference between α and β is the probable error.

A useful way of illustrating this is to say that if one observation is chosen at random out of a group, it is as likely as not that it will lie not further from the average than the probable error.

In the figure given at the end of the book, the probable errors are at the points P_1, P_2 for the curves A and C, and p_1, p_2 for the B.

By means of the approximation given in the last paragraph but two, the curve of error can be fitted to a series of observations by equating the probable error so determined to the value of $x = .47694$, and comparing the values of $F(x)$ with the ranged observations; but though this method is simple and rapid it is not the best.

By a suitable change of scale for ordinate and abscissa the equation given on p. 277 can be written $y = e^{-x^2}$, and this is the most general equation of the normal curve of error.

If $x = 0, y = 1$; hence the unit of ordinate is the number of cases at zero error; from the table on p. 281 above, the unit of abscissa is the probable error $\div .47694$. Since $\int_{-\infty}^{+\infty} e^{-x^2}.dx$ is shown in the integral calculus to be $\sqrt{\pi}$, the area contained between the curve and the axis of x is $\sqrt{\pi}$. If the equation is written $y = \frac{1}{\sqrt{\pi}} . e^{-x^2}$ its area is 1, that is, unit of area equals unit of probability (that is certainty) and the area contained by any two

* See Yule in *Statistical Journal*, 1896, p. 330.

ordinates, the curve and the axis of abscissa equals the probability of an occurrence between the errors represented by the abscissæ of those ordinates.*

If the curve is traced from either of the tables, it will be found that it changes from concavity to convexity on each side of the maximum ordinate, at such points as s_1 s_2 s_3 s_4 in the diagram at the end. If the unit of abscissa is taken as a concrete quantity, say, 1 inch, and the abscissa (OS_2) of this point of inflexion is ϵ, in the same units, then the equation of the curve is—

$y = \frac{1}{\sqrt{\pi}} . e^{-\frac{x^2}{2\epsilon^2}}$, for then $\frac{d^2y}{dx^2} = e^{-\frac{x^2}{2\epsilon^2}} \left(\frac{x^2}{\epsilon^4} - \frac{1}{\epsilon^2}\right) = 0$, if $x = \pm \epsilon$, and therefore the points $\left(\pm\epsilon, \frac{1}{\sqrt{\pi e}}\right)$ are points of inflexion.

Let $2\epsilon^2 = c^2 = \frac{1}{h^2}$, then the equation is $y = \frac{1}{\sqrt{\pi}} . e^{-\frac{x^2}{c^2}}$ or $y = \frac{1}{\sqrt{\pi}} e^{-x^2h^2}$

The area of this curve is $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} . e^{-\frac{x^2}{c^2}} . dx = c.$

Choose the unit of ordinate so that this area shall be unity. Then the equation is—

$$y = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}} = \frac{h}{\sqrt{\pi}} . e^{-h^2x^2}$$

c, which thus determines the unit of abscissæ of the curve, is called the *modulus*.

DETERMINATION OF THE MODULUS.

Suppose we have a series of observations which we know are selected from a group which conforms to the law of error, it is required to find, from the observations, the centre and modulus of the curve from which they would come with least improbability.

Let x_1, x_2, . . . x_n be the observed values. Let $\bar{x}$ be their arithmetic mean. Let δ_1, δ_2 . . . be the divergencies of the values from this mean.

---

* Hence chance of error between x and $x+dx$ is $\frac{1}{\sqrt{\pi}} . e^{-x^2} dx$.

Then $\delta_1 = x_1 - \bar{x}$, $\delta_2 = x_2 - \bar{x}$, . . . ; $\Sigma\delta = \Sigma_1^n x - n\bar{x} = o$, for $\bar{x} = \frac{\Sigma_1^n . x}{n}$.

Let the equation of the curve to which the observations belong be $y = \frac{1}{c\sqrt{\pi}} . e^{-\frac{(x-k)^2}{c^2}}$, where c and k have to be determined.

Let $y_1, y_2, \ldots y_r \ldots y_n$ be the values of y corresponding to $x_1, x_2, \ldots x_r \ldots, x_n$.

Then $y_r . dx$ is the chance that an object taken at random from the group conforming to the curve shall be between x_r and $x_r + dx$.*

Let P be the chance that the n given observations occur together.

Then $P = y_1 . y_2 \ldots y_n$

$$= \frac{1}{(c\sqrt{\pi})^n} e^{-\frac{(x_1-k)^2}{c^2}} \times e^{-\frac{(x_2-k)^2}{c^2}} \times \ldots \times \ldots\ldots \times e^{-\frac{(x_n-k)^2}{c^2}}$$

$$= \frac{1}{(c\sqrt{\pi})^n} . e^{-\frac{\Sigma_1^n (x-k)^2}{c^2}}$$

Now $\Sigma_1^n (x-k)^2 = \Sigma_1^n (\delta + \bar{x} - k)^2 = \Sigma_1^n \delta^2 + 2.(\bar{x} - k).\Sigma\delta + n.(\bar{x} - k)^2$ $= \Sigma\delta^2 + n(\bar{x} - k)^2$, since $\Sigma\delta = 0$.

Whatever value we assign to c, P will be greatest when the quantity $(\bar{x} - k)^2$ is least, that is when $k = \bar{x}$.

Giving k this value,

$$P = \frac{1}{(c\sqrt{\pi})^n} . e^{-\frac{\Sigma\delta^2}{c^2}}$$

$$\frac{dP}{dc} = (-nc^{-n-1} + 2c^{-n-3} . \Sigma\delta^2) . \pi^{-\frac{n}{2}} . e^{-\frac{\Sigma\delta^2}{c^2}}.$$

In order that P may be a maximum, $\frac{dP}{dc}$ must = 0.†

$$\text{Hence } c^2 = \frac{2\Sigma\delta^2}{n} \ddagger$$

Thus the curve required has its centre at the arithmetic

---

* When the magnitude of the observations is discontinuous, as in Quetelet's scale, no dx is necessary; but if the magnitude is continuous the probability of any defined error is zero, and the y is the chance of an error between infinitesimal limits.

† And $\frac{d^2P}{dc^2}$ be negative, which can be shown to be the case here.

‡ The proof here given is based on Merriman's *Method of Least Squares;* but it is suggested that his statement "for a given system of errors, it must be considered that the observations have been as precise as possible," § 65, is unnecessarily obscure.

mean of the observed values and modulus equal to $\sqrt{\frac{2\Sigma\delta^2}{n}}$ where the δ's are the differences between the observed values and their mean.*

c, so determined, is called the *modulus*. $h=\frac{1}{c}$ is called the *precision*. Professor Edgeworth proposes to call $c^2=\frac{2\Sigma\delta^2}{n}$ the *fluctuation*.†

It can be shown that half the curve $y=\frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}$ is included between the ordinates corresponding to $\pm r$, when $r=.47694\ c$, as found from the scale of precision.

It can be shown that the arithmetic average (η) of all the errors, considered positive,‡ is given by—

$$\eta=\frac{c}{\sqrt{\pi}}, \text{ whence } r=.8453\ \eta.$$

η is the abscissa of the centroid of the positive half of the curve.

η is more easily calculated from the observations than c, and can in some cases be used in its stead.§ η is called the *average error* or *mean of errors*.

In the table given on p. 281, the modulus is taken as 1; when $x=1$, F $(x)=.421\ldots$; that is .421.. of the curve lies between the ordinate corresponding to the abscissæ equal to the modulus and the central ordinate. When $x=2$, F $(x)=.4976\ldots$ Hence the chance of an observation showing a divergence from the mean on either side of more than twice the modulus $=.005\ldots$; the corresponding rank in Quetelet's table is 45.1.

ϵ used on p. 283, is now seen to be equal to $\sqrt{\frac{\Sigma\delta^2}{n}}$ and is called the *error of mean square*, or the *standard deviation*.¶

---

* See also p. 307, *infra*.

† *Stat. Journal:* Jubilee Number, p. 188; and p. 298, *infra*.

‡ $$\eta=\frac{\int_0^\infty xydx}{\int_0^\infty ydx}=\frac{-\frac{c}{2\sqrt{\pi}}\cdot\int_0^\infty e^{-z}.dz}{\frac{1}{2}}, \text{ where } z=-\frac{x^2}{c^2}, =\frac{c}{\sqrt{\pi}}$$

§ *Stat. Journal, loc. cit.*

¶ Both ϵ and η have been called *mean error*, and this term has become misleading.

In the diagram at the end OP_2, OP_1 are the probable errors, OM_1, OM_2 the moduli, OS_1, OS_2 the errors of mean square and OE_1, OE_2 the mean errors for the curves A and C. Op_1, Op_2, Om_1, Om_2, Oe_1, Oe_2, are corresponding quantities for the curve B. It is to be noticed that the line XOX_1 is an asymptote of each of the three curves drawn. In the figure OG_5 equals about twice the modulus. The ratio of the small area to the right of a vertical line through G_5 to the area of half the curve is the chance that twice the modulus shall be exceeded. The distance between OX_1 at three times the modulus and the curve is too small to be shown.

From the foregoing it will be seen that any curve of error, $y = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}$, can be obtained by projection from the same standard curve $y = e^{-x^2}$, just as any ellipse can be obtained by projection from a circle; but as ellipses differ from one another in virtue of different values of their eccentricity, so curves of error differ from one another in virtue of different values of their modulus. As we have seen, on any such curve there are certain definite points (the positions of the modulus, mean error, probable error, and error of mean square); if then we have the same units of abscissæ, such as 1 inch, for two sets of observations, these points will take different positions. If for one set the modulus is 2 inches, and for another 1 inch, .843 of the observations will be within 2 inches of the mean in the first case, and within 1 inch of the mean in the second. If we regard the observations as attempts to hit the mean and the divergencies as errors, the aiming in the second case is twice as precise as in the first. The precision thus defined is in inverse proportion to the modulus, and is therefore suitably measured by $h = \frac{1}{c}$. If the standard form of equation is adopted in both cases, the area of both curves will be unity, and their actual shapes those of C_1 C_2 C_3 and B_1 B_2 B_3.

The calculation of the precision of a set of observations does not require either of the tables, but is simply the evaluation of $h = \sqrt{\frac{n}{2\Sigma\delta^2}}$ from the observations themselves.

Another form of the equation in common use is $y = \frac{n}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}$ where n is the whole number of observations in question. The

area of this curve is n, so that unit area corresponds to one observation.

Compare now this form with that obtained from the limit of the binomial expansion, $y = \mathrm{P}.e^{-\frac{x^2}{2npq}}$

If $x = o$, $y = \mathrm{P}$; hence P is the maximum ordinate.

Adjusting the unit of ordinate so that the area of the curve is n, $y = \frac{n}{\sqrt{2npq\pi}} e^{-\frac{x^2}{2npq}}$

Hence $2npq = c^2$, and therefore $2np(1-p) = c^2$.

EXAMPLES.

I. The following figures, which are taken from Professor Westergaard's *Die Grundziige der Theorie der Statistik*, but treated in a manner different from his, will serve to illustrate the meaning of the formulæ, and show how to fit a curve to observations; limits of space prevent more elaborate examples.

BIRTHS IN DENMARK.

YEAR.	NUMBER.		Percentage Boys of Total.	Difference from Average.
	Total.	Boys.		
1860	54,797	28,308	51.66	+.23
1861	53,747	27,506	51.17	−.26
1862	53,011	27,300	51.50	+.07
1863	53,939	27,841	51.62	+.19
1864	52,884	27,334	51.68	+.25
1865	55,434	28,483	51.38	−.05
1866	57,353	29,747	51.87	+.44
1867	54,763	28,036	51.20	−.23
1868	56,546	28,985	51.26	−.17
1869	54,056	27,577	51.02	−.41
1870	56,472	29,144	51.60	+.17
1871	56,407	29,045	51.49	+.06
1872	57,274	29,462	51.44	+.01
1873	58,616	30,115	51.37	−.06
1874	59,324	30,594	51.57	+.14
1875	61,791	31,784	51.44	+.01
1876	63,967	32,912	51.45	+.02
1877	63,772	32,508	50.98	−.45
1878	63,144	32,505	51.48	+.05
1879	64,363	33,114	51.45	+.02
Average	57,583	...	51.43	...

Calculated modulus $\sqrt{\{2 \times 57583 \times .5143 \times .4857\}} = 169.61$ for a total of 57,583. Equivalent to .2945 . . . for a total of 100.

In the formula $\sqrt{2pqn}$, p the chance that any child born is a boy is taken as .5143, since 51.43 is the average percentage male births are of total births. $q = 1 - p = .4857$. The number of experiments n is taken to be the average number of births per year. Then $\sqrt{2pqn}$ is found to be 169.61. This is the modulus to apply to the whole number of births; but since this differs year by year it is convenient to reduce all numbers to percentages. The examples are then arranged as "between average + modulus and average — modulus," "between average + $\frac{9}{10}$ of modulus and average — $\frac{9}{10}$ of modulus," and so on; the first group (+.460) is taken so as to include the extreme.

Within.	Observed.	Calculated.	
		x.	$F(x) \times 2$.
51.44 ± .460	20	...	.972 of 20 = 19.4
51.44 ± .295	17	1.	.843 " = 16.8
51.44 ± .265	17	.9	.797 " = 15.9
51.44 ± .236	15	.8	.741 " = 14.8
51.44 ± .206	13	.7	.678 " = 13.6
51.44 ± .177	12	.6	.604 " = 12.1
51.44 ± .147	10	.5	.521 " = 10.4
51.44 ± .118	9	.4	.428 " = 8.6
51.44 ± .088	9	.3	.329 " = 6.6
51.44 ± .057	6	.2	.223 " = 4.5
51.44 ± .029	4	.1	.112 " = 2.2

The numbers to be expected from theory are given alongside; the fit is fairly close, and can be tested by the method described on p. 279.

The modulus calculated by the formula $\sqrt{\frac{2\Sigma\delta^2}{n}}$ is .305.

II. If a digit is taken at random, the chance that it will be less than 5 (0, 1, 2, 3, or 4) is $\frac{1}{2}$. If we take a book of logarithms and note the digits in the 7th decimal places in successive numbers, we shall have a practically random selection of digits. If we take groups of 50 numbers, the chances of 0, 1, 2 . . . 50 occurrences of digits less than 5 are given respectively by the terms of the expansion $(\frac{1}{2} + \frac{1}{2})^{50}$. This experiment was repeated 300 times and the results tested in accordance with both the

formulæ for the modulus, viz., $\sqrt{2pqn} = \sqrt{2 \times \frac{1}{2} \times \frac{1}{2} \times 50} = 5$, and $\sqrt{\frac{2\Sigma\delta^2}{n}}$, which was found to be 4.75.

The values of the other standard errors were both found directly and deduced from c, with very fair correspondence between the two methods.

NUMBER OF OCCURRENCES OF 0, 1, 2, 3, OR 4 IN THE 7TH DECIMAL PLACES OF GROUPS OF 50 LOGARITHMS.

										Averages of Lines.
29	19	25	25	22	28	16	23	22	27	23.6
24	28	30	22	20	27	24	24	27	22	24.8
27	26	28	21	21	22	22	27	25	25	24.4
25	28	21	23	22	23	27	27	25	25	24.6
28	23	26	23	22	29	28	25	23	26	25.3
24	22	22	19	26	24	26	28	20	25	23.6
24	23	27	29	26	21	26	31	23	27	25.7
26	26	30	25	25	24	29	25	21	27	25.8
30	24	25	27	24	30	24	28	24	30	26.6
26	21	21	22	31	28	26	26	26	33	26.0
19	25	26	34	21	28	21	29	19	23	24.5
29	26	19	29	24	27	25	25	24	22	25.0
24	27	21	23	25	21	26	28	25	27	24.7
25	28	29	30	28	27	28	23	25	26	26.9
30	30	18	22	24	23	25	27	25	31	25.5
25	26	27	21	23	25	24	20	25	22	23.8
28	24	20	18	25	19	25	30	29	25	24.3
22	27	24	28	22	20	23	25	26	26	24.3
16	27	28	27	23	20	29	26	20	24	24.0
26	28	28	23	21	24	25	21	14	28	23.8
25	24	21	24	24	21	24	30	25	26	24.4
28	32	17	23	29	24	22	33	29	29	26.6
25	26	26	27	22	20	24	26	24	24	24.4
28	26	25	25	29	25	24	22	26	25	25.5
22	30	22	27	25	27	27	27	28	21	25.6
35	26	23	26	31	28	26	22	22	29	26.8
26	20	28	23	22	28	24	23	30	16	24.0
28	26	25	31	27	28	22	26	30	19	26.2
26	27	27	18	25	24	27	22	30	30	25.6
26	25	24	17	27	32	25	21	23	30	25.0
Averages of Columns.										
25.9	25.7	24.4	24.4	24.5	24.9	24.8	25.7	24.5	25.7	

1.	2.	3. Errors.	4. Squares of Error.	Product of Col. 2 and Col. 3.	Product of Col. 2 and Col. 4.
	Times.				
14	occurs 1	− 11.04	121.88	− 11.04	121.88
16	" 3	− 9.04	81.72	− 27.12	245.16
17	" 2	− 8.04	64.64	− 16.08	129.28
18	" 3	− 7.04	49.56	− 21.12	148.68
19	" 7	− 6.04	36.48	− 42.28	255.36
20	" 9	− 5.04	25.40	− 45.36	228.60
21	" 18	− 4.04	16.32	− 72.72	293.76
22	" 26	− 3.04	9.24	− 79.04	240.24
23	" 21	− 2.04	4.16	− 42.84	87.36
24	" 32	− 1.04	1.08	− 33.28	34.56
25	" 42	− .04	0	− 1.68	0
26	" 36	+ .96	.92	+ 34.56	33.12
27	" 30	+ 1.96	3.84	+ 58.80	115.20
28	" 28	+ 2.96	8.76	+ 82.88	245.28
29	" 15	+ 3.96	15.68	+ 59.40	235.20
30	" 16	+ 4.96	24.60	+ 79.36	393.60
31	" 5	+ 5.96	35.52	+ 29.80	177.60
32	" 2	+ 6.96	48.44	+ 13.92	96.88
33	" 2	+ 7.96	63.36	+ 15.92	126.72
34	" 1	+ 8.96	80.28	+ 8.96	80.28
35	" 1	+ 9.96	99.20	+ 9.96	99.20
...	300	...	Sum of Errors, all considered positive	785.12	3387.96 Sum of Squares.

Average, 25.04. Median, 25. Quartiles, 27, 23. Hence probable error is approximately 2.

$$\text{Mean error} = \frac{785.12}{300} = 2.617.$$

$$\text{Error of mean square} = \sqrt{\frac{3387.96}{300}} = 3.36 = \epsilon.$$

$$\text{Modulus} = \epsilon \times \sqrt{2} = 4.75 = c.$$

Also probable error $= .4769\, c = 2.265 = r$, and mean error $= r \div .8453 = 2.68$, nearly the values obtained directly.

The following table compares the distribution with that of the normal curve, by a method differing from those previously used.

Divergence of 1 from average corresponds to $x = \frac{1.}{4.75} = .21.$

Between Average and	Observed.					x.	$2F(x)$.
Average	Above Average.		Below Average.		Total.		
± 1	36	+	42	=	78	.21	.117 of 600 = 70
± 2	66	+	74	=	140	.42	.224 " = 134
± 3	94	+	95	=	189	.63	.314 " = 188
± 4	109	+	121	=	230	.84	.383 " = 230
± 5	125	+	139	=	264	1.05	.431 " = 259
± 6	130	+	148	=	278	1.26	.463 " = 277
± 7	132	+	155	=	287	1.47	.481 " = 289
± 8	134	+	158	=	292	1.68	.491 " = 295
± 9	135	+	160	=	295	1.89	.496 " = 298
±10	136	+	163	=	299	2.10	.499 " = 299
±11	136	+	163	=	299	2.31	.499 " = 300
±12	136	+	164	=	300	2.52	.499 " = 300

The fit is close. The symmetry is spoilt by the great number 42 at 25 occurrences, just below the average.

The line $L_1L_2 \ldots L_{22}$ on the diagram of the curve of error at the end of the book shows these numbers. The moduli are made to correspond, which defines the abscissæ, and the scale of ordinates is then decided by making the areas of the two figures equal, but this was not done exactly.

Summary of Terms.

For convenience of reference the principal quantities connected with the curve of error are collected below.

If we take $y = \frac{1}{c\sqrt{\pi}} . e^{-\frac{x^2}{c^2}}$ as the equation of the curve, c which determines the unit of abscissæ is called the *modulus*.

If δ_1 δ_2.... are the differences between the observations and their arithmetic average, c should be taken as $\sqrt{\frac{2\Sigma\delta^2}{n}}$ or $\sqrt{\frac{2\Sigma\delta^2}{n-1}}$, where n is the number of observations. (See pages 285 and 307.)

If the curve can be also determined as the limit of an assigned binomial expansion $(p+q)^n$, then c may be taken as equal to $\sqrt{2pqn}$. (See page 287.)

$\frac{2\Sigma\delta^2}{n}$ or $\frac{2\Sigma\delta^2}{n-1}$ is called the *fluctuation*, and equals c^2. (See pages 285 and 307.)

$h, = \frac{1}{c}$, is the *precision.*

The square root of the average of the squares of the δ's, $= \sqrt{\frac{\Sigma\delta^2}{n}}$, is called the *error of mean square*, or the *standard deviation*, and is represented by the letters ϵ or σ.

Hence $c = \epsilon . \sqrt{2}$.

The arithmetic average of all the δ's, all reckoned as positive, is called the *average error.* It is equal to the distance of the centre of gravity of half the curve from the central ordinate. It is represented by the letter η, and $\eta = \frac{c}{\sqrt{\pi}} = .5641896c$.

The *probable error* is half the distance between the quartiles of the observations. The ordinates through the points whose abscissæ are the probable errors bisect the two symmetrical halves of the curve. The probable error is represented by the letter r, and $r = .4769363c$. This is generally written $r = \rho c$, where $\rho = .4769363$.

c, h, ϵ, η and r can all be calculated directly from the observations. In general the values so calculated will not satisfy the above numerical relations exactly; the correspondence depends on the closeness of the fit of the observations to the curve.

LIBRARY OF THE BOMBAY BRANCH OF THE ROYAL ASIATIC SOCIETY

SECTION III.—TO WHAT GROUPS DOES LAW OF ERROR APPLY?

Returning to our discussion on the relation between the laws of probability and the numerical facts of actual experience, let us consider the meaning of such phrases as "a rare occurrence," "an improbable event," "a run of luck," "a lucky man," and similar expressions which show that some events are regarded as ordinary, others as extraordinary. On this subject there is a great deal of popular confusion; thus the *Spectator* opens its columns to people who write about extraordinary coincidences, *e.g.*, that on 3rd March in two successive years two persons of the same name died at the same age in neighbouring villages; and recently the concurrence of the two names Arthur and Mallory in a dispatch was instanced as remarkable. Now, *à priori*, these two names are just as likely to be mentioned together as any other two borne by equal numbers of persons. If out of n persons, p bear the one and q the other, the chance that the first two names given in an assigned place will be these is about $\frac{2p}{n} \times \frac{q}{n}$; but the chance that they will occur together in the newspapers in a given week is much greater, viz., $\frac{2pq}{n^2} \times N$, where N is the number of pairs of names in conjunction in all the columns of the press together. Going a step further, consider the number of pairs of names that, when placed together, would recall some event of historic or other interest, and suppose this to be M; then the chance that some such coincidence should arise in a given week is $\frac{2pq}{n^2} \times N \times M$, if we suppose for the sake of argument that p and q are the same for all the pairs concerned. From these remarks it will be seen that before we can speak of an event as extraordinary, we must define the time, place, circumstances, and nature of such events. Further, suppose we decide to regard an event as unusual if the chance of its occurrence thus defined is less than $\frac{1}{r}$, where r is a large number, it is easily seen that we may expect the improbable, to speak paradoxically; for great though r may be, the

The meaning of luck.

number of events which come under our cognizance is also great; and we may therefore expect to find on an average one improbable event for every r we notice; hence it is possible for a weekly newspaper with the help of the widely-extended search for sensations of its intelligence department to supply us week by week with its quantum of horrors. Another aspect of the same subject will be seen when we deal with the permanence and regularity of certain small numbers in Section IV.

The idea of rarity.

The rarity of an event is often unconsciously determined by a mental forecast of its occurrence. If I take four cards out of a well-shuffled pack and find them to be in succession, ace, king, queen, knave of hearts, I should feel surprised; not because these four cards are less likely to come than any other four assigned cards whatever, but because I have certain associations with them in that they form a sequence which is valuable in certain games, and are the highest cards of a suit; there are noted in my mind unconsciously many groups of four cards of such special significance. If there are s such groups, the chance that one of them will occur is $\frac{s}{{}_{52}C_4}$, if we do not, and $\frac{s}{{}_{52}P_4}$, if we do, regard the order of their occurrence.

The real difference between a rare and a common event is, however, independent of any mental process or prejudices. If I place 8 coins one after another in front of me, it is no more unlikely that I shall get 8 heads than that I shall get any other *assigned order* of heads and tails, say *h t t h t h t t;* the chance of either is $\frac{1}{2^8}$; but it is much more unlikely that I shall get 8 heads than that I shall get 5 tails and 3 heads *without regard to the order* in which they come; for out of $(2^8 =)$ 256 possible arrangements, only 1 gives 8 heads, but $\left(\frac{8!}{3!5!} =\right)56$ give 5 tails and 3 heads.

The greatness of the improbability dealt with by the law of error.

Apply this argument to our hypothesis as to great numbers. Suppose the population to be composed of males and females in equal numbers, and that 1,000 persons are selected on some system quite unconnected with sex. Out of $2^{1000} (= 10^{301})$ possible selections (differing from another only in arrangement in order of sexes) only 1 gives 1,000 males, but $\frac{1000!}{500!500!} (= 10^{297} \times 27)$ arrange-

ments gives 500 of each sex, independently of the order. The chance of the first occurrence is $\frac{1}{10^{301}}$,* of the second about 1 in 37. In statistics we are concerned with the totals, depending only on the *combinations* of the items, not on their order (the permutations); and occurrences of the numbers near the average (500, 499, 501, &c.) are separately and much more conjointly very much more probable than occurrences of the numbers far from it. The vast improbability of very great divergence can be seen by a numerical study of the curve of error (see p. 281).

Hence the theorems relating to great numbers rest on a very much firmer basis than they would if divergence was due to that sort of coincidence which produces a so-called rare event.

A "run of luck," good or bad, may be regarded as a succession of improbable events, and is a more scientific expression than a "rare event" as commonly understood. Of a great number of events, deals of cards, investments, bets, and so on, very many will give normal results, average success at cards, normal returns to investments and so on; very few will give abnormal winnings. The chance of abnormal success in one venture being p, a small fraction, the chance of a succession of n successes is p^n, very much smaller when n is at all large. It is in the phrase "lucky man" that the error is introduced. One who has benefited by the occurrence of a rare event may reasonably be called lucky, and the number of lucky men will be roughly proportional to the number of fortunate rare events; but when a succession of events, say three, each of probability $\frac{1}{100}$, and conjointly of probability $\frac{1}{1000000}$, or a broken succession (*e.g.*, *ppqpp* of which chance would be $\frac{1}{20000000}$) has taken place in one man's favour, the imagination loses the logic of the case, and supposes an overruling law, and marks out that particular man as not subject to the law of probability: one is apt to expect that the next event will also be a success, and to be further confirmed in this opinion by paying attention to

A common fallacy.

* Chance of 1,000 m. *or* 1,000 f. is twice this.

the one instance when the sixth event is a success, and neglecting the ninety and nine when it is a failure. Other people are biassed in the opposite direction, and have distinctly too great an expectation of a counterbalancing tendency, a long run of failures till the average is restored. It is thus correct to speak of a man having been lucky, but tempting Nemesis to speak of him as a lucky man. It is a mere truism to say that, unless a success or failure have some causal influence over future successes or failures (as when a good stroke at a game steadies the nerves for another), the probability of each future event is totally unaffected by what has gone before.

Let us return now to the method of deducing the chance p, and the index n, used in the expansion $(p+q)^n$, from records such as the death-rate. Notice first that the deduction of p (the chance to be applied to each individual to find the varying degrees of probability of the possible totals) from the numbers, implies some hypothesis as to the genesis of these numbers, the very theorem which we wish it to illustrate; for suppose that in the records of 20 years we find 600,000 deaths in a stationary population of 1,000,000, we assume that this is the most probable number which a chance régime would give, and since the most probable number is the total $\times p$, we deduce that $p = \frac{600000}{1000000} \times \frac{1}{20} = \frac{3}{100}$; but here we are making some undefined assumption about the occurrence of events similar to that defined by the curve of error. If we actually assumed the law of error, we can calculate how far the value of p so estimated may be expected to differ from its true value. This accounts in part for the divergence between the calculated grouping and the fact.

Statistical coefficients.

Again, there is great difficulty in determining the number n, the number of persons to whom the chance of the occurrence of a particular event is applied, and we should further notice that in many cases, in particular in concrete measurements, such as height and normal length of life, we have no information whatever as to n, which in this case is the number of causes which may add or subtract undefined units from height or age; and we are often equally in doubt when dealing with great numbers, *e.g.*, with the total value of imports. In such cases we should have to deduce both p

and n from the records of results; and indeed it is simpler to fall back on other methods of deducing the law of error than the present one of regarding it as the limit of the binomial expansion, determining the modulus without any assumption as to the number of independent causes. Hence in a great many instances we cannot expect to find close conformity to a predetermined curve $(p+q)^n$. Similarly we can deduce from the laws of gravitation and motion that a planet's orbit must be an ellipse, but cannot determine the eccentricity of this ellipse except by observation.

Reason of apparent non-correspondence of fact and theory.

A far-reaching cause of the apparent discrepancy between fact and theory is, however, of a different kind. The theory applies to experiments performed under unchanging conditions; if we are drawing differently coloured balls from a bag containing a great number, all the external circumstances must be unchanged, and the only variation that which comes from the so-to-say regulated randomness of the forces which decide shuffling and drawing. Now in human affairs, when we consider a series of death-rates or any other rates distributed in time, we are dealing with a constantly changing environment of social and sanitary habits, within which the apparently random forces that decide death are acting; and these external changes may affect the interaction of the random forces, just as a change in barometric pressure may affect the molecular forces of a rigid body. Such effects cannot be foretold or calculated; we may expect that improvements in sanitation will diminish the death-rate, but some detail may increase it; vaccination may diminish small-pox, but increase the liability to some other disease. To such reasons as these should be assigned the non-correspondence to the law of error of great numbers distributed in time. When the element of time is eliminated by a process of random averaging the correspondence is closer. Great numbers distributed in space are exempt from this disturbing cause and might be expected to show closer correspondence; for instance the birth-rates in a number of districts might be expected to conform more closely than rates for one place for a series of years; but it is very difficult to obtain sufficiently homogeneous figures distributed in space, though Prof. Lexis gives some instances of this kind.*

* *Massenerscheinung*, p. 66.

Physiological and anthropometrical measurements, such as the heights of 10,000 children of the same age, are not affected by these difficulties, and should show close correspondence with the theoretical distribution; and it is not surprising that the ratio of the number of male to the number of female births, depending as it does on hidden causes not easily influenced by the progress of civilisation, should show that remarkable consilience with the law of error, which has so often been remarked. Finally, the occurrence of sequences and groups of numbers, such as those obtained from logarithmic tables, being absolutely independent of changes in time or space, naturally show complete agreement with theory.

All these considerations make the application of the law of error to actual measurements a very delicate operation, and it may appear that the cases where agreement is close are so few as to make the whole body of theory useless; but this is an unscientific view to take. The general process of applied science is to frame hypotheses as nearly consistent with the facts as is possible without such complications as will prevent their use, and then apply to the idealized case the corrections which the actual cases necessitate. This process has led to the best results in physical science. In the problems dealt with by the law of error, it will be found that many deductions from the idealized cases hold also when applied to the only partially corresponding records of great numbers; just as, in mechanics, many theorems relating to smooth bodies can be applied unchanged to rough bodies. For instance, the "fluctuation" of non-corresponding figures can be calculated by the formula $\frac{2\Sigma\delta^2}{n}$; and the accuracy of an average of random samples of quantities not grouped according to the curve of error varies as the square root of the number of samples taken.*

The use of the law of error.

From this discussion we may gather that we can seldom tell *à priori* whether the law of error will or will not apply to a given series of figures. This must be determined by experiment for each new class of records; but when we have found correspondence in many series of a class (as is the case in measurement of heights) we may proceed with confidence to apply the law to other similar series or groups.

An important distinction is drawn by Prof. Lexis† and em-

* See p. 308.

† *Ibid.*, p. 28.

phasised by Prof. Edgeworth * between two classes of figures to which the laws of great numbers apply. The first, called by Lexis concrete, contains such quantities as height measurements of a great number of persons, and normal length of life, where a definite mean or type seems to be normal and other measurements to be variations from this type. In these cases it is not easy to connect the facts which correspond with the exponential curve $y = e^{-x^2}$, where x is the divergence from the type, with our deduction from the limit (n infinite) of $(p+q)^n$. Suppose, however, that height is determined by n forces, each capable of adding or of subtracting 1 unit, say 1 millimetre, from normal height, and that the chance that each shall act is p; then the divergencies obtained in a number of individuals should be distributed according to the coefficients in this expansion.

Concrete measurements and great numbers.

The other class, called by Lexis *combinational*, to which the discussion in Section II. above more directly applies, contains those totals which are the sum of a great number of items (persons, deaths, births, &c.), for the existence of each of which a definite chance, p, can be assigned *à priori*. The numbers may then be expected (subject to the disturbing causes already discussed) to be grouped in accordance with the curve $y = \frac{n}{\sqrt{2\pi pqn}} e^{-\frac{x^2}{2pqn}}$, where n is the total numbers of persons to whom the chance p applies. In such cases p is the arithmetic average of $p_1 p_2 \ldots p_n$, where $p_1 z$, $p_2 z$, $\ldots$ $p_n z$ are the numbers of events which are found respectively in n series each of z observations. p is the "probability coefficient" of the event, and $p_1, p_2, \ldots p_n$ should conform to a curve with modulus $\frac{\sqrt{2p(1-p)}}{z}$. On the other hand, if c is calculated from the formula—

$$c^2 = \frac{2}{n-1}\left\{\left(p-p_1\right)^2 + \left(p-p_2\right)^2 + \ldots + \left(p-p_n\right)^2\right\} = \frac{2\Sigma(d^2)}{n-1}$$

we are treating the p's as "concrete" quantities, and obtain a second value for the modulus.

If $\sqrt{\frac{2p(1-p)}{z}} - \sqrt{\frac{2\Sigma(d^2)}{n-1}} = 0$, the distribution of the coefficients $p_1, p_2, \ldots p_n$ is normal, which is not often the case.

* Jubilee Volume of *Statistical Journal*, p. 191.

If this quantity <0, the coefficients are grouped more closely together than the theory of error leads us to expect, and there is some evidence that a force preventing divergence has been called into play.

More generally this quantity is >0, the coefficients are more divergent than in accordance with the theory of error, and some disturbing forces have acted.

SECTION IV.—THE PERMANENCE OF CERTAIN SMALL NUMBERS.

The binomial expansion and small numbers.

A remarkable side-light is thrown on our general argument by the actual permanence of small numbers. Little attention has been given to this phenomenon, but it is a very striking fact that if among a great number of items there are a few which present some particular feature, it will be found that this small number is seldom much exceeded and seldom entirely vanishes.

The following numerical example shows that this may be expected theoretically, and an examination of the successive terms of $(p+q)^n$ when p is very small and q nearly equal to 1 will show the same phenomenon more generally.

CONSTANCY OF SMALL NUMBERS.

$$\left(\frac{1}{1001}+\frac{1000}{1001}\right)^{4000} = \frac{1}{1001^{4000}}\left(1000+1\right)^{4000}$$

1st term $= 1000^{4000} \div 1001^{4000}$.

$$\log\left(\frac{1000}{1001}\right)^{4000} = 4000\,(3-3.0004341) = -1.7364 = \bar{2}.2636 = \log .018348.$$

No occurrences	Chance of			
		1st term	$= a$, suppose,	$= .0183$
1	,,	2nd ,,	$= 4000 \times 1000^{3999} \div 1001^{4000} = 4a$	$= .0734$
2	,,	3rd ,,	$= \frac{4000 \times 3999}{1.2} \times 1000^{3998} \div 1001^{4000}$	
			$= \frac{4000 \times (4000-1)}{1.2} \times 1000^{3998} \div 1001^{4000}$	
			$= 8 \times 1000^{4000} \div 1001^{4000}$, correct to 1 in $4000 = 8a$	$= .1467$
3	,,	4th ,,	$= \frac{4000 \times 3999 \times 3998}{1.2.3} \times 1000^{3997} \div 1001^{4000}$	
			$= \frac{4000 \times (4000^2 - 3 \times 4000)}{1.2.3} \times 1000^{3997} \div 1001^{4000}$	
			$= \frac{4^3}{1.2.3} \times 1000^{4000} \div 1001^{4000}$, less 3 in 4000 approx.	$= .1956$
4	,,	5th ,,	$= \frac{4^4}{\underline{4}}\,a$, less 6 in 4000 approx.	$= .1954$
5	,,	6th ,,	$= \frac{4^5}{\underline{5}}\,a$, less 10 in 4000 ,,	$= .1562$
6	,,	7th ,,	$= \frac{4^6}{\underline{6}}\,a$, less 15 in 4000 ,,	$= .1040$
7	,,	8th ,,	$= \frac{4^7}{\underline{7}}\,a$, less 22 in 4000 ,,	$= .0593$
8	,,	9th ,,	$= \frac{4^8}{\underline{8}}\,a$, less 30 in 4000 ,,	$= .0296$
			10th	$.0131$
			11th	$.0052$
			12th	$.0019$
			13th	$.0006$
			14th	$.0002$

Terms 15 to 4001 together only occur about 1 in 10000.

Chance of 3, 4, 5, or 6 occurrences $= \frac{7}{10}$ approx.

To take an actual example:—Out of some 530,000 deaths annually from all causes the following are the numbers from splenic fever in the years 1875-1894:—

5, 4, 10, 14, 12, 18, 9, 15, 8, 18, 11, 11, 11, 12, 7, 4, 3, 6, 7, 10.
Average 10.

$$\text{Here } p = \frac{10}{530000}; \quad \therefore q = \frac{52999}{53000}.$$

n is doubtful, and may be taken to be the total number of deaths or the total population; but it will be found that the following numbers are unaffected, whichever number we adopt.

The successive terms in the expansion $(p + q)^{530000}$ are given in the second column.

Chance of		Number of occurrences to be expected in 20 years.		Number observed.	
0 deaths is	.000045	0	0	0	1
1 „	.00045	0		0	
2 „	.00225	0		0	
3 „	.0075	.2		1	
4 „	.0185	.4	6	2	
5 „	.037	.7		1	
6 „	.061	1.2		1	
7 „	.087	1.7		2	
8 „	.11	2.2	7	1	
9 „	.12	2.4		1	
10 „	.12	2.4		2	
11 „	.11	2.2		3	
12 „	.09	1.8	4	2	
13 „	.07	1.4		0	
14 „	.05	1.		1	
15 „	.03	.6		1	
16 „	.02	.4	2	0	
17 „	.01	.2		0	
18 „	.00	0		2	
More than 18	0	0		0	

Considering the small number of years taken, and the indefiniteness of many of the death returns, the general consilience between the last two columns is satisfactory; while the general principle that small numbers show a certain constancy is well exemplified. Specialists in all professions, from the doctor who treats only one obscure disease of the ear, to the dealer in curiosities, make their livelihood dependent on this permanence of small numbers.

The regular occurrence of accidents and of improbable events in general furnishes other examples of the same sort.

Note.—Since writing this section my attention has been called to a treatise by Dr Bortkewitsch, *Das Gesetz der Kleinen Zahlen*, Leipzig, 1898, where the close agreement of the records of accidents and other occasional events to the binomial expansion is dealt with in a more exhaustive and analytical manner.

SECTION V.—EXTENSION OF THE LAW OF ERROR AND APPLICATIONS.

Generalized statement of law of error.

We have only shown so far that great numbers fluctuate about their mean in accordance with the law of error on the assumption that for the existence or non-existence of each particular unit there is the same numerical chance p. We can, however, prove by elementary methods that the same distribution is reached under many other circumstances, and at the same time make several important deductions.

Suppose that a quantity whose mean value is H is determined by the action of a great number of causes; let the causes produce deviations ϵ_1, ϵ_2, . . . which are connected with η, the corresponding deviation of H, by the equation $\eta = a_1\epsilon_1 + a_2\epsilon_2 + +$, where a_1, a_2 . . . are constants. If each of the deviations ϵ_1, ϵ_2 . . . can be of various magnitudes, the curves which show the probabilities of the occurrence of these magnitudes are called "curves of frequency," or "facility curves." If the curves of frequency are normal curves of error, the chance of the occurrence of the deviation ϵ_1, ϵ_2 . . . are proportional to $e^{-\frac{\epsilon_1^2}{c_1^2}}$, $e^{-\frac{\epsilon_2^2}{c_2^2}}$ . . . where c_1, c_2 . . . are the moduli of these curves. The following proof holds when these assumptions are justified; but the resulting theorems hold (1) when the ϵ's belong to any curves of frequency such that their limits are narrow, while the number of ϵ's is great, and the limits of each of the ϵ's is small compared with η, and none of the a's are predominant; (2) where η is any function of $(a_1\epsilon_1 + a_2\epsilon_2 + \ . \ . \ .)$, such that $\eta = a_1\epsilon_1 + +$ is a first approximation.*

The equation to the normal curve can also be deduced directly from other considerations, when we are dealing with any quantity liable to continuous small independent variations.†

We will now show that when the assumptions are limited

* Adapted from a paper by Professor Edgeworth in the *London and Edinburgh Philosophical Magazine*, vol. 34, 1892, p. 429.

† See, for instance, Chauvenet's *Astronomy*, vol. ii., Appendix; and Merriman's *Method of Least Squares*, chap. ii.

as above that η belongs to a normal curve of error with modulus $\sqrt{(a_1^2c_1^2+a_2^2c_2^2+\ \ )}$.

CASE I.—If H is determined by two causes only, $\eta = a_1\epsilon_1 + a_2\epsilon_2$. Probability that the deviations ϵ_1, ϵ_2 concur $= C.e^{-\frac{\epsilon_1^2}{c_1^2}} \times e^{-\frac{\epsilon_2^2}{c_2^2}}$ (where C is a constant) = (eliminating ϵ_2)

$$C.e^{-\left\{\frac{\epsilon_1^2}{c_1^2}+\frac{(\eta-a_1\epsilon_1)^2}{a_2^2c_2^2}\right\}} = C.e^{-\frac{\eta^2}{(a_1^2c_1^2+a_2^2c_2^2)}} \times e^{-\frac{a_1^2c_1^2+a_2^2c_2^2}{a_2^2c_1^2c_2^2}\cdot\left(\epsilon_1-\frac{a_1c_1^2\eta}{a_1^2c_1^2+a_2^2c_2^2}\right)^2}$$

This quantity is the probability that a deviation ϵ_1 occurs with a deviation η. Giving ϵ_1 all its possible values in turn, we find that the probability of a deviation η is

$$C.e^{-\frac{\eta^2}{(a_1^2c_1^2+a_2^2c_2^2)}} \times \sum_{\epsilon_1=-\infty}^{\epsilon_1=\infty} e^{-\frac{a_2^2c_1^2+a_2^2c_2^2}{a_2^2c_1^2c_2^2}\left(\epsilon_1-\frac{a_1c_1^2\eta}{a_1^2c_2^2+a_2^2c_2^2}\right)^2}$$

Now the quantity included in the summation is the whole of a normal curve of error, which has been shifted through a horizontal distance $\frac{a_1c_1^2\eta}{a_1^2c_1^2+a_2^2c_2^2}$, and its value depends only on the constants a_1, a_2, c_1, c_2, and is independent of η and ϵ_1.

Hence probability of a deviation $\eta = e^{-\frac{\eta^2}{a_1^2c_1^2+a_2^2c_2^2}} \times$ constant, and η belongs to a normal curve of error with modulus $\sqrt{a_1^2c_1^2+a_2^2c_2^2}$.

CASE II.—If H is determined by three causes, $\eta = a_1\epsilon_1 + a_2\epsilon_2 + a_3\epsilon_3$, write η_1 for $a_1\epsilon_1 + a_2\epsilon_2$, then $\eta = \eta_1 + a_3\epsilon_3$.

By theorem of Case I., modulus for η_1 is $\sqrt{a_1^2c_1^2 + a_2^2c_2^2}$. and modulus for η is

$$\sqrt{\left[\left\{\sqrt{a_1^2c_1^2+a_2^2c_2^2}\right\}^2+a_3^2c_3^2\right]} = \sqrt{(a_1^2c_1^2+a_2^2c_2^2+a_3^2c_3^2)}.$$

Similarly the theorem can be extended to any number of causes.

Corollaries.

1. Let $\bar{x}$ be the weighted average of the quantities $x_1, x_2 \ldots$ with weights $a_1, a_2 \ldots$, so that $\bar{x} = \frac{\Sigma ax}{\Sigma a}$;

suppose that the weights are known accurately, but that $x_1 = x_1^1 + \epsilon_1$, $x_2 = x_2^1 + \epsilon_2, \ldots$ where $x_1^1, x_2^1 \ldots$ are correct values, and $\epsilon_1, \epsilon_2 \ldots$ errors belonging to normal curves with moduli $c_1, c_2 \ldots$ Then if $\bar{x}^1$ be the correct value of $\bar{x}_1$

$$\bar{x} = \frac{\Sigma a(x^1+\epsilon)}{\Sigma a} = \bar{x}^1 + \frac{\Sigma a\epsilon}{\Sigma a}.$$

Hence the modulus of $\bar{x}$ is $\frac{\sqrt{(a_1^2c_1^2 + a_2^2c_2^2 + \ldots)}}{a_1 + a_2 + \ldots}$, for each term in the above theorem can be divided by the constant $a_1 + a_2 + \ldots$*

2. Putting $a_1 = a_2 = a_3 =$, and $c_1 = c_2 = c_3 = \ = c$, we find as before that modulus of an unweighted average of n quantities, conforming to a curve with modulus c, is $\frac{\sqrt{n} \cdot c}{n} = \frac{c}{\sqrt{n}}$.

3. If H is the difference between two quantities whose mean values are the same, and moduli c_1 and c_2,

η, the deviation in H, $= \epsilon_1 - \epsilon_2$; modulus for $\eta = \sqrt{c_1^2 + c_2^2}$.

4. If the two quantities are the averages of n_1 and n_2 quantities with moduli v_1, v_2, then by corollary 2, $c_1 = \frac{v_1}{\sqrt{n_1}}$, $c_2 = \frac{v_2}{\sqrt{n_2}}$, and modulus for difference between the quantities is by corollary 3, $\sqrt{\frac{v_1^2}{n_1} + \frac{v_2^2}{n_2}}$.

5. In particular, the modulus for the difference between two quantities from one group, modulus c, is $c\sqrt{2}$.

Corollaries 3, 4, and 5 can be proved directly by the method of Case I.

Precision of an Average.

Precision of an average.

The second corollary, that the precision of an average is proportional to the square root of the number of terms it contains, is so important that an independent proof may be offered, starting from different assumptions.

Suppose a great number (m) of observations to be made of a single unknown quantity, *e.g.*, the declination of a star. Let r be the "probable error" of a single observation, h the precision of the group, v the true, $v + d_1$, $v + d_2$, . . . $v + d_m$ the observed values of the quantity. Let x_0 be the arithmetic mean of the m observed values, and let $x_0 = v + \delta$. Then δ is the error of the arithmetic mean.

Let δ_1, δ_2, . . δ_m be the differences between the observed values and x_0, so that $v + d_1 = x_0 + \delta_1 = v + \delta + \delta_1$; $d_1 = \delta + \delta_1$, $d_2 = \delta + \delta_2$, &c.; also $mx_0 = \Sigma_1^m (x_0 + \delta_1) = mx_0 + \Sigma_1^m \delta_1$; and $\Sigma_1^m \delta_1 = 0$.

* This should be compared with pp. 204-214, *supra*.

Let P_1 be the probability that this set of observations concur. Then

$$P_1 = \frac{h}{\sqrt{\pi}} e^{-h^2 d_1^2} \times \frac{h}{\sqrt{\pi}} e^{-h^2 d_2^2} \times \text{to } m \text{ products} = \frac{h^m}{\pi^{\frac{m}{2}}} e^{-h^2 (d_1^2 + d_2^2 + \ldots)}$$

$$= \frac{h^m}{\pi^{\frac{m}{2}}} e^{-h^2 \Sigma_1^m (\delta + \delta_1)^2} = \frac{h^m}{\pi^{\frac{m}{2}}} e^{-h^2 \left\{ m\delta^2 + 2\delta\, \Sigma_1^m (\delta_1) + \Sigma_1^m (\delta_1^2) \right\}}$$

$$= \frac{h^m}{\pi^{\frac{m}{2}}} e^{-h^2 m\delta^2} \times e^{-h^2 \Sigma_1^m (\delta_1^2)}, \text{ since } \Sigma_1^m (\delta_1) = 0.$$

P_1 is the probability that the observed values yield an error δ in their arithmetic mean. Let P_0 be the probability that the observed values yield no error in their arithmetic mean, then

$$P_0 = \frac{h^m}{\pi^{\frac{m}{2}}} e^{-h^2 \Sigma_1^m (\delta_1^2)}$$

$$P_1 = P_0 \times e^{-h^2 m\delta^2}$$

Hence the arithmetic mean belongs to a curve of error whose precision is $\sqrt{h^2 m} = h\sqrt{m}$, and therefore its probable error is $\frac{r}{h\sqrt{m}}$.

If the errors $d_1, d_2 \ldots$ occur $p_1, p_2 \ldots$ times respectively in the observations, while $p_1 + p_2 + \ldots + p_m = n$, the foregoing argument is unaffected, and the precision of the mean is $h\sqrt{n}$, that is $h\sqrt{(p_1 + p_2 + \ldots + p_m)}$.

Care is needed to distinguish the hypotheses on which this formula, and the former formula $\frac{\sqrt{\Sigma a^2}}{\Sigma a}$ connecting weights, depend.

A corresponding result may be obtained directly from the limit of the binomial expansion. If an experiment, for whose success the chance is p, is performed n times, the most probable number of successes is the nearest integer to pn, and the modulus for the various numbers is $\sqrt{2pqn}$. The modulus for the average of the n experiments is therefore $\frac{\sqrt{2pqn}}{n} = \frac{\sqrt{2pq}}{\sqrt{n}}$; that is, the precision is proportional to $\sqrt{n}$.

We can now obtain a formula for the modulus of a series of observations in a form often given. On p. 285 it is shown that if $\delta_1, \delta_2 \ldots \delta_n$ are divergencies from their average of a series of observations, and if these divergencies conform to a law of error with modulus c, then c should be taken as $\sqrt{\frac{2\Sigma\delta^2}{n}}$

and the centre of the curve at the average, for maximum probability. Now the average from which these divergencies are measured, conforms to a curve modulus $\frac{c_1}{\sqrt{n}}$, where c_1 is the modulus of the divergencies measured from their *true value*, not from their arithmetic mean;

then, if Δ is the divergence from the true value, modulus c_1,
δ is the divergence from the arithmetical mean, modulus c,
d is the divergence of the arithmetic mean from the true value, modulus $\frac{c_1}{\sqrt{n}}$,

$$\Delta = \delta + d$$

and $c_1^2 = c^2 + \frac{c_1^2}{n}$, from page 304.

$$\text{Hence } c_1^2 = \frac{nc^2}{n-1} = \frac{2\Sigma\delta^2}{n-1}.$$

Since n is large these quantities are very nearly equal; and it is not worth while here to discuss their relative merits; the latter quantity $\sqrt{\frac{2\Sigma d^2}{n-1}}$ is generally used.*

As an example of this greater precision of averages, take the averages given on p. 289, each of 30 numbers, which range on a normal curve modulus 5; these averages are 25.9, 25.7, 24.4, 24.4, 24.5, 24.9, 24.8, 25.7, 24.5, 25.7. General average 25.05; differences, .85, .65, −.65, −.65, −.55, −.15, −.25, .65, −.55, 65; $\sqrt{\frac{2\Sigma d^2}{9}} = .89$. The modulus for such groups is by theory $\frac{5}{\sqrt{30}} = .91 \ldots$

From the same page we may find the following 30 averages, each of 10 numbers:—23.6, 24.8, 24.4, 24.6, 25.3, 23.6, 25.7, 25.8, 26.6, 26.0, 24.5, 25.0, 24.7, 26.9, 25.5, 23.8, 24.3, 24.3, 24.0, 23.8, 24.4, 26.6, 24.4, 25.5, 25.6, 26.8, 24.0, 26.2, 25.6, 25.0.

The probable error for these is by theory .47 of $\frac{5}{\sqrt{10}} = .642$; and between the limits $25.04 \pm .64$ we actually find 15 out of the 30 averages, while 6 are below the lower, and 7 above the higher limit.

---

* *Vide* article by Prof. Edgeworth, *Camb. Phil. Soc. Trans.*, 1885.

The modulus for the whole 300 is $\sqrt{\frac{2 \times \frac{1}{2} \times \frac{1}{2} \times 50}{300}} = .29$ and the probable error .14; the average for an infinite number would be 25; for the 300 selected it is 25.043, that is well within the probable error.

Examples of this kind could be multiplied indefinitely.

SAMPLES.

Sampling.

The bearing of this principle on the method of sampling is very important. Our experience on most subjects is derived, not by examining all the existing examples, but by noting a few which come in our way. A man of specialized experience is one who has seen and analyzed mentally many cognate phenomena. It needs no proof that the more samples taken, the more accurate will be the judgment formed about the group of which they are samples. Very many business transactions are decided by such an examination. Now we have seen that the precision of the average shown by samples of quantities which satisfy the normal law of error is inversely proportional to the square root of their number; but there are three further questions to consider—(α) Whether this rule applies to samples of quantities which do not conform to the law of error, that is, which would not be obtained from a normal distribution without great improbability; (β) how we are to measure the precision of either the original group of which we have samples or of our samples; (γ) whether we can learn anything more about the original group besides its average.

α. On referring back to page 303, it will be seen that the averages of samples of, say, m quantities drawn at random from a large group whose distribution is not normal, will, if m is large in relation to the fluctuation of the original group, satisfy the law of error. The reason, apart from the mathematical analysis of this, is clearer from the following illustration: if we have records of a quantity, which fluctuates in accordance with the normal law about an average which changes slowly year by year, our measurements will not conform to the normal law; but if we select four years at random again and again, we shall eliminate the influence of time, and our samples will tend to conform. Readers may experiment on the annual birth-rates to illustrate this.

The following numbers are the death-rates per 10,000 in London registration districts, arranged in order of magnitude.—

70	100	113	120	130	141	150	160	170	181	191	204	230	252	323
70	107	115	121	130	141	150	163	177	183	194	205	236	252	329
80	108	115	121	131	141	150	164	178	183	198	210	237	255	329
92	108	115	123	132	142	151	166		185	198	211	238	264	404
	109	116	123	132	144	151	167		188		220		264	448
		117	124	132	144	152	167				222		266	475
		118	125	133	144	152	168				222		276	505
		118	126	135	144	153					223		284	622
			126	136	145	154					228		286	625
			127	137	145	155								1,408
			128	138	147	155								
				138	148	156								
				139	149	156								
				139		157								
						158								
						158								

These numbers clearly do not conform to the normal curve. We will omit 1,408 as being so far from the others as to be in a class by itself and select at random samples of 4, 18 times. Their averages are 174, 222, 226½, 221, 129, 150, 181½, 193, 300, 133, 216, 178, 167, 169½, 183, 150, 227, 164. Average, 188; modulus, 57.4. These fit a curve of error closely, thus—

Within 5 of average	Observed.	Calculated from Table on p. 281.
	2	1.7
,, 6½ ,,	3	2.3
,, 10 ,,	4	3.5
,, 14 ,,	5	5.5
,, 18½ ,,	6	6.3
,, 21 ,,	7	7.1
,, 24 ,,	8	8.0
,, 28 ,,	9	9.2
,, 33 ,,	10	10.5
,, 34 ,,	11	10.7
,, 38 ,,	12	11.3
,, 38½ ,,	14	11.8
,, 39 ,,	15	11.9
,, 55 ,,	16	14.8
,, 59 ,,	17	16.7
,, 112 ,,	18	18.0

Thus the theorem is confirmed in a very unpromising case.

β. To determine the precision of the average of our samples, two methods are open. The first consists in finding the modulus $c=\sqrt{\frac{2\Sigma d^2}{n-1}}$ of all the quantities chosen; then if the quantities conform to a normal curve the modulus of their average is $\frac{c}{\sqrt{n}}=\sqrt{\frac{2\Sigma d^2}{n(n-1)}}$, and the precision is $\frac{\sqrt{n}}{c}$; if the quantities do

not conform this formula still gives the best measure of the precision, but it may be well to confirm it by the second method. This method is to break up the n samples into $\frac{n}{m}$ smaller groups each of m, and see if the averages of these groups are such as would come from a normal distribution; if they do not, increase m; if they show signs of normal grouping in a curve of modulus c, before we have come to the limiting value of m, then we may expect that the larger sample of n things belongs to a normal curve, whose modulus is $\frac{c_1}{\sqrt{\frac{n}{m}}}$, which may be expected to be equal to $\frac{c}{\sqrt{n}}$. If we do not get conformity with the largest value of m we can take, we have no guarantee that n is large enough to eliminate the abnormality of the original figures.

The following statistics of wages give a practical application of this principle.

Numerical example. In the period 1834-45 inquiries were made in the Scotch villages as to the day wages of agricultural labourers.

The resulting figures for the Lowlands may be tabulated as follows:—

Numbers at 13d.	13½d.	14d.	15d.	16d.	16½d.	17d.	17½d.	18d.	18½d.
5	3	2	8	12	6	24	3	39	3

Numbers at 19d.	20d.	21d.	22d.	23d.	23½d.	24d.	24½d.	25d.	27d.
27	26	27	15	1	1	4	1	2	2

Average, 18.8d.; modulus, 3.62d. $= c$.

CORRESPONDENCE WITH LAW OF ERROR.

Limits.	Normal	Observed: Above Average.		Below Average.		Total.
$18.8 \pm \frac{1}{5}c$	46	27	+	3	=	30
$\frac{2}{5}c$	90	53	+	45	=	98
$\frac{3}{5}c$	127	53	+	69	=	122
$\frac{4}{5}c$	156	80	+	87	=	167
c	178	95	+	87	=	182
$\frac{6}{5}c$	192	96	+	95	=	191
$\frac{7}{5}c$	201	97	+	97	=	194
$\frac{8}{5}c$	206	102	+	100	=	202
$2c$	210	104	+	105	=	209

When we divide the returns into 50 samples of 4 we get modulus for their averages 1.8; 25 samples of 8 give modulus 1.14; 40 samples of 5 give modulus 1.57; 20 samples of 10 give modulus 1.19.

The c for the original samples may be found from any of these; the results are—

Modulus of original samples				3.62
,,	,,	calculated from the groups of	4,	$1.8 \times \sqrt{4} = 3.6$
,,	,,	,,	8,	$1.14 \times \sqrt{8} = 3.2$
,,	,,	,,	5,	$1.57 \times \sqrt{5} = 3.5$
,,	,,	,,	10,	$1.19 \times \sqrt{10} = 3.8$

This is a close consilience with theory. We will adopt 3.6 as the value of c, then the modulus of the average of the 211 original samples is $\frac{3.6}{\sqrt{211}}$, its precision $\frac{\sqrt{211}}{3.6}$, and its probable error .47 of $\frac{3.6}{\sqrt{211}} = .12\ldots$, or $\frac{1}{8}$ of a penny.

We should verify that the samples conform to the law of error: the following shows the comparison for the samples of 4:—

Limits.	Normal.	Observed: Above Average.		Observed: Below Average.		Total.
18.8 ± $\frac{1}{5}$ of modulus (1.8)	11	6	+	7	=	13
$\frac{2}{5}$,,	21	7	+	11	=	18
$\frac{3}{5}$,,	30	14	+	17	=	31
$\frac{4}{5}$,,	37	14	+	23	=	37
modulus	42	18	+	24	=	42
$\frac{6}{5}$ of modulus	45	20	+	26	=	46
$\frac{7}{5}$,,	48	23	+	26	=	49
$\frac{8}{5}$,,	49	23	+	26	=	49
$\frac{9}{5}$,,	49	23	+	27	=	50
2 modulus	50	23	+	27	=	50

This resemblance is as close as the argument requires.

γ. If our first samples conform to the law of error we know with reasonable certainty the average and the distribution of the original quantities—namely, that they conform to a normal curve with approximately the same average and modulus as our samples. The general average and the sample average differ in accordance with a law of error, modulus $\frac{c}{\sqrt{n}}$, where c is modulus for samples and n their number.

If our first samples do not conform, it is still probable that their curve of frequency has a resemblance to that of the original quantities. If the fraction p_1 of the original quantities lay between assigned limits a_0 and a_1, then the number to be expected between those limits in n samples is decided by the expansion of $(p_1+q_1)^n$, where $p_1+q_1=1$; the most probable number is the integer nearest $p_1.n$, and the modulus is $\sqrt{2p_1q_1n}$; similarly if p_2, q_2 bear a similar relation to a_1, a_2, the most probable number selected between these limits is $p_2.n$, and modulus $\sqrt{2p_2q_2n}$, and so on. Thus a similar distribution may be expected, and each part of it has a precision varying jointly as the square root of the whole number taken and the quantity $\sqrt{p(1-p)}$; thus the larger the number taken the greater will be the resemblance, and [since $\sqrt{p_1(1-p_1)} > \sqrt{p_2(1-p_2)}$ when $p_1 > p_2$ and $p_1+p_2 < 1$] the larger the altitude of the area in the curve of frequency corresponding to given limits the greater its precision. The errors of the various divisions are not, however, entirely independent of one another. This is, of course, in strict accordance with the common sense of the question.

The following examples of school ages illustrate part of this argument. In a school containing 257 boys of varying ages, where the dispersion was not likely to be normal, 48 were selected at random and their ages written down.

Numerical example.

The modulus of the 48 samples is 43.2 months; their average 13 years 10 months; their distribution as follows:—

	Normal.	Observed: Above Average.		Below Average.		Total.
Average $\pm \frac{1}{6}$ modulus	11	3	+	6	=	9
$\frac{2}{6}$,,	21	10	+	10	=	20
$\frac{3}{6}$,,	29	12	+	16	=	32
$\frac{4}{6}$,,	36	13	+	21	=	34
modulus	40	16	+	22	=	38
$\frac{5}{6}$,,	44	19	+	24	=	43
$\frac{7}{6}$,,	46	20	+	26	=	46
$\frac{8}{6}$,,	47	22	+	26	=	48

The observations are not grouped symmetrically nor in close agreement with the normal distribution.

When we take the average of random samples we do not find close relation to the normal curve till the number of samples becomes too small to work with. To continue the

process we assume that 48 is a large enough number of items to neutralize the want of symmetry in the figures. The average of the whole group is as likely as not to be within the limits 13 years 10 months' $\pm \frac{43}{\sqrt{48}} \times .47$ months, that is between 13 years 7 months and 14 years 1 month.

Again the quartiles in our samples are at 18 months above and 2 years below the average; the quartiles in the original group may be expected to be within the same distances with probable errors $\sqrt{2 \times \frac{1}{4} \times \frac{3}{4} \text{ of } 48} = 4.2$ months, since the chance that any quantity shall be between the average and the lower quartile is $\frac{1}{4}$.

From a census of the whole school it was found that all these conditions were fulfilled; the average was 14 years; the quartiles were unfortunately not kept; but 58 boys out of the 257 were stated to be over 15 years 9 months, from which it is highly probable that the upper quartile was within the given limits, 15 years 6 months $\pm$ 4 months; and 54 were below 11 years 10 months, which places the lower quartile also well within the limits.

Corollary 4, relating to the modulus of a difference, is most useful in comparing two groups selected as having certain qualities. Thus Professor Edgeworth* discusses whether an ascertained difference of 2 inches between the average heights of a large number of criminals and that of the general population is significant; and finding that the modulus for the difference between two random groups is only 0.08, holds that there is a cause of the difference in the method of selection; that is, that criminality and low stature are found together. We might apply the same principle to the investigation of the existence of a period in any figures; for if the modulus of the figures was c, the modulus for the difference between the averages of two random samples of 20 months each would be $c\sqrt{\frac{1}{20}+\frac{1}{20}} = \frac{c}{\sqrt{10}}$; if the difference between the averages of the figures for 20 Decembers and 20 Junes was 3 times this quantity, the existence of a period would be established. For instance, in the percentage of ironfounders unemployed monthly from 1855 to 1874† the modulus for single months

Precision of a difference.

* *Statistical Journal*, Jubilee Number, *ibid.*

† See p. 179, *supra.*

is about 30, and for the difference between the averages of two groups one of 20 and the other of 240 is therefore $30.\sqrt{\frac{1}{20}+\frac{1}{240}}=7$; but the average of the 20 Decembers is about 29 °/₀ above the general average, a significant difference; and the average of the 20 Augusts is about 19 °/₀ below, a divergence smaller than before, but still significant; the difference between the Decembers and Augusts, namely, 48, is to be compared with the modulus $30\times\sqrt{\frac{1}{20}+\frac{1}{20}}=9$, and is therefore significant.

A final example may be given which brings into relation many of these theorems. The following were the recorded times for "The Oaks" from 1850 to 1899; we will discuss whether there has been a significant increase of speed, or some change in the conditions of the race, or whether the fluctuations are due to minor causes varying year by year.

General example.

	min.	sec.		min.	sec.		min.	sec.		min.	sec.		min.	sec.
1850	2	56	1860	2	56	1870	2	52	1880	2	49	1890	2	40½
1851	2	52	1861	2	44	1871	2	51	1881	2	46	1891	2	54⅔
1852	3	0	1862	2	49	1872	2	52	1882	2	49	1892	2	43½
1853	2	52	1863	2	54	1873	2	50¾	1883	2	53	1893	2	44⅔
1854	3	0	1864	2	47	1874	2	48½	1884	2	49	1894	2	50
1855	2	58	1865	2	51	1875	2	49½	1885	2	43⅔	1895	2	48⅚
1856	3	4	1866	2	53	1876	2	50	1886	2	54⅔	1896	2	45⅔
1857	2	50	1867	2	54	1877	2	54½	1887	2	50⅔	1897	2	45
1858	2	53½	1868	2	47½	1878	2	54	1888	2	42⅚	1898	2	45⅔
1859	2	55	1869	2	59	1879	3	2	1889	2	45	1899	2	44
Ten yearly averages	2	56		2	51½		2	52½		2	48		2	47

These figures fit fairly closely a normal curve of error with modulus 7.43 secs., average 2 min. 50.87 secs. The modulus for the difference between two is therefore $7.43\sqrt{1+1}=10.48$ secs The greatest difference between consecutive years is 14 secs., between 1856 and 1857; this is not sufficiently far beyond the modulus to make it uncommon; hence there is no proof of any sudden change in arrangements having taken place between two races. The difference between the times for years early in the period and those later is sometimes as much as 20 secs. The modulus for the difference between the averages for two periods of 10 years is $7.43\sqrt{\frac{1}{10}+\frac{1}{10}}=3.3$. The difference between the averages for

1850-59 and 1890-99 is 9 secs., which is significant; that between 1850-59 and 1880-89 is also significant. The odds against such a difference as that between the average times of 1850-59 and 1860-69 are only 13 to 1, not very significant. Hence we find that some cause was at work which gradually quickened the race between the fifties and the eighties.

This method can be applied to the criticism of such serial figures as birth, death, and marriage rates, imports, exports, and so on. With a periodic series the method can be used first for establishing the period, and then for investigation of the figures found when the periodicity is eliminated. With a symptomatic * curve, the method can be used for measuring the symptomatic tendency, and then for studying the short-period fluctuations. For a series which has no symptom and no period, the method is at once applicable for finding what divergencies are significant, and for forecasting and interpolating numbers. Without some machinery of calculation of this kind we are unable to get beyond vague and general impressions of the existence of a change; † but if we take care that the conditions of the calculation are satisfied, we can by the method now developed make a definite statement quite independent of personal bias, such as "either an event has happened, so improbable as to be outside the range of human experience, or the decrease shown in the series of figures in question is due to some significant change in the system of causes which produce them."

Application to series of different classes.

---

* See p. 240, *supra.*

† We can take an intermediate step by noticing in the above table that in nine cases out of ten the times in the decade 1880-9 are less than the times thirty years earlier; the chance that so great an agreement in the direction of the change (irrespective of its magnitude) should come in a random selection is $\frac{2}{1}\frac{2}{0}\frac{2}{2}\frac{2}{4}$ or .0215; the chance as calculated above is .0006. See Edgeworth in *Jubilee Volume of Statistical Journal*, pp. 213-217.

SECTION VI.—THE THEORY OF CORRELATION.

Causal connection.

It is never easy to establish the existence of a causal connection between two phenomena or series of phenomena; but a great deal of light can often be thrown by the application of algebraic probability. We have already dealt with some cases in point; we have shown how to find whether an event is due to a special cause, or whether it naturally arises from the variation of existing causes; we have shown how to measure the significance of the difference between two quantities or two averages; and further, we have investigated such problems as the influence of the seasons.* In many large groups of phenomena we can apply a more refined and more certain method, which it is our object to introduce in this section. When two quantities are so related that the fluctuations in one are in sympathy with the fluctuations of the other, so that an increase or decrease of one is found in connection with an increase or decrease (or inversely) of the other, and the greater the magnitude of the changes in the one, the greater the magnitude of the changes in the other, the quantities are said to be *correlated. Correlation* is a quantity which can be measured numerically; and its measurement has been the subject of much recent mathematical investigation.

The correlation surface.

Let two variable quantities X, Y be subject to variations x, y, which are due to a multitude of individually unimportant causes, producing fluctuations $e_1, e_2 \ldots \epsilon_1 \epsilon_2 \ldots$ so that the x's are connected with the e's and the y's with the ϵ's by the equations,

$$x = a_1 e_1 + a_2 e_2 + \quad + a_n e_n$$
$$y = b_1 \epsilon_1 + b_2 \epsilon_2 + \quad + b_n \epsilon_n, \text{ where } a_1, a_2 \ldots b_1, b_2 \ldots \text{ are constants.}$$

Then x and y conform to normal curves of error, whose moduli we will call c_1, c_2.

The rest of our investigation, which is based closely on Professor Karl Pearson's paper on "Regression, Heredity, and Panmixia,"† proceeds on the assumption that the e's and ϵ's

* See p. 186, *supra.*

† *Transactions of the Royal Society*, vol. 187 (1896), A. 253-318.

conform to normal curves of error, which is not, however, the most general condition.

Let individual values of X and Y, x and y respectively, be grouped in pairs, as measurements of two quantities at the same date, or of two parts of the same organism, or in any other way. If x and y are quite independent, none of the causes producing them are common to both, and the e's are independent of the ϵ's in the above equations. Then z, the chance of divergencies x and y concurring $= e^{-\frac{x^2}{c_1^2}} \times e^{-\frac{y^2}{c_2^2}} \times$ (a constant).

For *any* one value of x, the quantities y are grouped about the mean value Y_x in accordance with the normal curve c_2; and similarly for any one value of y.

The above equation may be written $z = C.e^{-\left(\frac{x^2}{c_1^2}+\frac{y^2}{c_2^2}\right)}$. If we give z any definite value k, the x's and y's which have jointly the probability k, are connected by the equation

$$\frac{x^2}{c_1^2} + \frac{y^2}{c_2^2} = \log \frac{C}{k}$$

which is the equation of an ellipse having its principal axes coincident with the axes of measurement of x and y, if we suppose x and y measured on two horizontal lines perpendicular to one another. Let z be measured vertically; then in the surface given by the equation connecting z, x, y all the horizontal sections are similar ellipses, whose projections on a horizontal plane are concentric and similarly situated,* while all the vertical sections are normal curves of error.†

This is the surface of no correlation.

If, on the other hand, any of the e's coincide with any of the ϵ's, it may be shown that a new term is introduced in the equation between z, x, and y, which becomes

$$z = \frac{n}{\pi c_1 c_2} \cdot \frac{1}{\sqrt{1-r^2}} e^{-\left(\frac{x^2}{c_1^2} - r\cdot\frac{2xy}{c_1c_2} + \frac{y^2}{c_2^2}\right)\cdot\frac{1}{1-r^2}}$$

* For a diagram of this projection and for a general discussion of correlation on the same lines as this chapter, but more advanced and complete, see Mr Udny Yule's paper on *The Theory of Correlation* in the *Statistical Journal*, Dec. 1897.

† For the section by a plane $y = mx + n$ is $z = Ce^{-\left(\frac{x^2}{c_1^2} + \frac{\overline{mx+n}^2}{c_2^2}\right)}$, which may be written $z = e^{-A(x+B)^2} \times D$ where A, B, and D are constants.

where n is the number of pairs of observations and r is a quantity we have still to decide; this is the general equation of the normal correlation surface. The horizontal sections, obtained by giving z different constant values, are now of the form

$$\frac{x^2}{c_1^2} - 2r.\frac{xy}{c_1c_2} + \frac{y^2}{c_2^2} = l, \text{ a quantity independent of } x \text{ and } y.$$

The projections of the horizontal sections are still concentric, similar and similarly situated ellipses, but their principal axes are now inclined to the axes of x and y. The vertical sections are still normal curves of error with various centres; in particular the frequencies of the values of y found in conjunction with x_1, a particular value of x, are given by the equation

$$z = e^{-\left(\frac{x_1^2}{c_1^2} - r.\frac{2x_1y}{c_1c_2} + \frac{y^2}{c_2^2}\right)\frac{1}{1-r^2}} \times \text{constant.}$$

$$= e^{-\frac{1}{c_2^2(1-r^2)}\left(y - r.\frac{c_2}{c_1}x_1\right)^2} \times \text{constant.}$$

This is a normal curve of error with its centre at $r.\frac{c_2}{c_1}x_1$.

Thus the mean value of y *corresponding to* x_1, *a given value of* x, *is* $r.\frac{c_2}{c_1}x_1$.

These mean values all lie on the line $\frac{y}{c_2} = r.\frac{x}{c_1}$.

Similarly the mean values of x, corresponding to given values of y, lie on the line $\frac{x}{c_1} = r.\frac{y}{c_2}$.

r is called the *coefficient of correlation.* If r is positive, for every given value of x, the mean value of the corresponding y's is positive and a definite fraction of x; if r is negative, the correlation is said to be negative, and for every given value of x, the mean value of the corresponding y's is a definite negative fraction of x.

The coefficient of correlation.

To determine the value of r, we must observe that this single quantity determines the shape of the whole surface, when N, c_1, c_2 are given, just as the modulus determines the shape of the curve of error. We decided the best value of the modulus* by considering from what curve of error the observed values would arise with least improbability. Professor Pearson finds the value

* *See* p. 283, *supra.*

of r by considering from what distribution of z, x, and y (*i.e.*, from what surface of correlation) the observed pairs would arise with least improbability; r is thus found to be $\frac{\Sigma(xy)}{n\sigma_1\sigma_2}$ or $\frac{1}{n}\cdot\sum\left(\frac{x}{\sigma_1}\cdot\frac{y}{\sigma_2}\right)$, the summation being extended over all pairs of x, y, where σ_1, σ_2 are the errors of mean square of the x's and y's respectively, and hence $\sigma_1 = \frac{c_1}{\sqrt{2}}$, $\sigma_2 = \frac{c_2}{\sqrt{2}}$.

But with other values of r the observed pairs might have been obtained with greater or less improbability, and these values are distributed in accordance with a normal curve of error whose probable error is $.67\frac{1-r^2}{\sqrt{n}}$; * that is, when from all the possible correlation surfaces, which might have resulted in the observed values, those whose correlation coefficients are within the limits $r \pm .67\frac{1-r^2}{\sqrt{n}}$ are selected, the sum of their probabilities is $\frac{1}{2}$.

It will be useful to examine the limits of the possible values of r.

r always lies between $+1$ and -1.

For $n^2\sigma_1^2\sigma_2^2 - (\Sigma xy)^2 = \Sigma x^2.\Sigma y^2 - (\Sigma xy)^2$, since $\sigma_1 = \sqrt{\frac{\Sigma x^2}{n}}$, $\sigma_2 = \sqrt{\frac{\Sigma y^2}{n}}$

$$= (\lambda_1^2 y_1^2 + \lambda_2^2 y_2^2 + +)(y_1^2 + y_2^2 + +) - (\lambda_1 y_1^2 + \lambda_2 y_2^2 + +)^2,$$

$$\left(\text{where } (x_1y_1)(x_2y_2) \ldots \text{ are pairs of observations, and } \frac{x_1}{y_1} = \lambda_1, \frac{x_2}{y_2} = \lambda_2 \ldots\right)$$

$$= y_1^2y_2^2(\lambda_1^2 + \lambda_2^2 - 2\lambda_1\lambda_2) + +$$

$$= y_1^2y_2^2(\lambda_1 - \lambda_2)^2 + +$$

which is zero if $\lambda_1 = \lambda_2 = \lambda_3 = =$, but otherwise positive.

Hence $n^2\sigma_1^2\sigma_2^2 - (\Sigma xy)^2$ is positive,

$$1 > \left(\frac{\Sigma xy}{n\sigma_1\sigma_2}\right)^2$$

$$1 > r^2$$

and r is between $+1$ and -1, except when $\lambda_1 = \lambda_2 = \lambda_3 =$; in this case $r = \pm 1$, and the correlation is said to be perfect, positively or negatively.

---

* *See* Pearson, *loc. cit.*, p. 226; Yule, *loc. cit.*, p. 847; and Pearson, *Proceedings of Royal Society*, Oct. 1897. By a similar line of reasoning the probable error of c as determined on p. 283 is found to be $.477\frac{c}{\sqrt{n}}$.

It may be noticed that on *à priori* grounds without any mathematical investigation the formula $\frac{1}{n}\left(\frac{x_1}{\sigma_1}.\frac{y_1}{\sigma_2}+\frac{x_2}{\sigma_1}.\frac{y_2}{\sigma_2}+\ +\right)$ gives a good measure of correlation. For if there is positive correlation, whenever we have a positive value of x we may expect a positive value of y, and whenever we have a negative value of x we may expect a negative value of y, and each such term increases the coefficient; while, if there is no correlation, for any value of x occurring several times, we may expect positive and negative values of y which on the whole give a very small sum. Meanwhile the denominators at once bring the deviations into relation with the mean deviations, and prevent the whole coefficient becoming greater than unity.

We see then that r measures the correspondence between deviations from their means of the two series of observations.

The measurement of correlation. If the deviations are in exactly the same ratio for all pairs, the correlation is perfect, and $r = 1$; while r tends to zero when for a given deviation in one of the series we have excess and defect with equal frequency in the other.

r serves as a measure of any statement involving two qualifying adjectives, which can be measured numerically, such as "tall men have tall sons," "wet springs bring dry summers," "short hours go with high wages."

When r is not greater than its probable error $.67\frac{1-r^2}{\sqrt{n}}$ we have no evidence that there is any correlation, for the observed phenomena might easily arise from totally unconnected causes; but when r is greater than, say, 6 times its probable error, we may be practically certain that the phenomena are not independent of each other, for the chance that the observed results would be obtained from unconnected causes is practically zero.

The calculation of r is quite simple, and if we can assume normal dispersion, so that the probable error in a series is equal to .67 of the error of mean square,* can be performed very rapidly. In the following tables the correlation between the prices of wheat, foreign trade, and the marriage-rate, already discussed by the help of the graphic method, is investigated.

* Hence σ = about $\frac{3}{4}$ of distance between quartiles.

EXAMPLES OF CORRELATION.*

Marriage-Rate and Price of Wheat.

1845-64.

Years.	Marriage-Rate.	Differences.	Price of Wheat. s.	d.	Differences.	Products of Differences.	Imports and Exports £ mln.
1845	17.2	+ .44	50	10	− 20	− 8	143
1846	17.2	+ .44	54	8	+ 26	+ 11	131
1847	15.8	− .96	69	9	+ 207	− 199	142
1848	15.9	− .86	50	6	− 24	+ 21	142
1849	16.2	− .56	44	3	− 99	+ 56	162
1850	17.2	+ .44	40	3	− 147	− 65	172
1851	17.2	+ .44	38	6	− 168	− 74	185
1852	17.4	+ .64	40	9	− 145	− 92	187
1853	17.9	+ 1.14	53	3	+ 9	+ 10	222
1854	17.2	+ .44	72	5	+ 239	+ 105	249
1855	16.2	− .56	74	8	+ 262	− 147	239
1856	16.7	− .06	69	2	+ 200	− 12	288
1857	16.5	− .26	56	4	+ 46	− 12	310
1858	16.0	− .76	44	2	− 114	+ 87	281
1859	17.0	+ .24	43	9	− 119	− 29	309
1860	17.1	+ .34	53	3	+ 9	+ 3	346
1861	16.3	− .46	55	4	+ 34	− 16	342
1862	16.1	− .66	55	5	+ 35	− 23	340
1863	16.8	+ .64	44	9	− 93	− 4	396
1864	17.2	+ .44	40	2	− 148	− 63	435
Av.	16.76		Av. 52	6		$\Sigma xy = -461$	

Correlation between marriage-rate and the price of wheat—

$\sigma_1 = .580 \qquad \sigma_2 = 133$

$$r = \frac{-461}{20 \times 133 \times .58} = -.30$$

Probable error of $r = -.14$

Correlation between marriage-rate and imports and exports—

$\sigma_1 = .580 \qquad \sigma_3 = 90$

$\Sigma xy = 8$

$r = +.007$

Probable error of $r = .15$

* For a complete examination of the particular question here used for illustration, see Mr Hooker's paper in the *Statistical Journal*, September 1891.

1875-94.							
1875	16.7	+ 1.53	45	2	+ 88	+ 135	597
1876	16.5	+ 1.33	46	2	+ 100	+ 133	576
1877	15.7	+ .53	56	9	+ 227	+ 120	593
1878	15.2	+ .03	46	5	+ 103	+ 3	562
1879	14.4	− .77	43	10	+ 72	− 55	554
1880	14.9	− .27	44	4	+ 78	− 21	634
1881	15.1	− .07	45	4	+ 90	− 6	631
1882	15.5	+ .33	45	1	+ 87	+ 29	654
1883	15.5	+ .33	41	7	+ 45	+ 15	667
1884	15.1	− .17	35	8	− 26	+ 4	623
1885	14.5	− .67	32	10	− 60	+ 40	584
1886	14.2	− .97	31	0	− 82	+ 80	563
1887	14.4	− .77	32	6	− 64	+ 49	584
1888	14.4	− .77	31	10	− 72	+ 55	622
1889	15.0	− .17	29	9	− 97	+ 16	676
1890	15.5	+ .33	31	11	− 72	− 23	684
1891	15.6	+ .43	37	0	− 10	− 4	683
1892	15.4	+ .23	30	3	− 91	− 21	651
1893	14.7	− .47	26	4	− 138	+ 65	623
1894	15.1	− .07	22	10	− 180	+ 13	624
	Av. 15.17		Av. 37	10		$\Sigma xy = 627$	

Correlation between marriage-rate and the price of wheat—

$\sigma_1 = .651 \quad \sigma_2 = 102$

$\Sigma xy = 627$

[Or distance between quartiles = .9, whence $\sigma_1 = .67$]

$$r = \frac{627}{20 \times 102 \times .651} = +.47$$

Probable error .1

Correlation between marriage-rate and imports and exports—

$\sigma_1 = .651 \quad \sigma_3 = 41$

$r = +.25$

Probable error of $r = .14$

Hence there was slight negative correlation between the marriage-rate and price of wheat before 1864, that is, the marriage-rate fell when wheat rose; but since 1864 there is better evidence that the marriage-rate rises when wheat rises. The marriage-rate and foreign trade were quite uncorrelated before 1864, and show only slight correlation at more recent dates; the odds against the correspondence between the observed figures, since 1875, arising without causal connection are only about 4 to 1, if we assume that the figures for each year are independent of the next.

The Galtonic method.

An earlier method of estimating correlation, introduced by Mr Galton,* is very useful for a rapid survey of two groups of figures. As a simple example adequately illustrating the method, we will take two series

* See *Proceedings of the Royal Society*, 1886, vol. xi., *Family Likeness in Stature*.

CORRELATION OF DAILY MAXIMA AND MINIMA OF TEMPERATURE IN 1898.

MINIMA

MAXIMA: Not below	but below	23° – 27°	27° – 31°	31° – 35°	35° – 39°	39° – 43°	43° – 47°	47° – 51°	51° – 55°	55° – 59°	59° – 63°	63° & above	Number of maxima falling within limits	Medians of corresponding minima
35° & below 40°		3	2	9									14	32°
40° & below 45°		2	7	12	7	6							34	33°
45° & below 50°		1	4	17	13	11	5						51	36°
50° & below 55°			4	5	12	12	11	8					52	40½°
55° & below 60°			1	1	3	15	16	15	2	1			54	44½°
60° & below 65°				1	3	3	16	15	10	2			50	47°
65° & below 70°					1	7	3	9	5	3			28	48°
70° & below 75°							2	7	10	5	2		26	53°
75° & below 80°							3	5	7	9	2		26	54°
80° & below 85°									1	10	5		16	58°
85° & above											4	3	7	63°
Number of minima falling within limits		6	18	45	39	54	56	59	35	30	13	3	TOTAL 358	
Medians of corresponding maxima		40°	45°	45°	50°	54°	59°	62°	70°	77°	82°	87°		

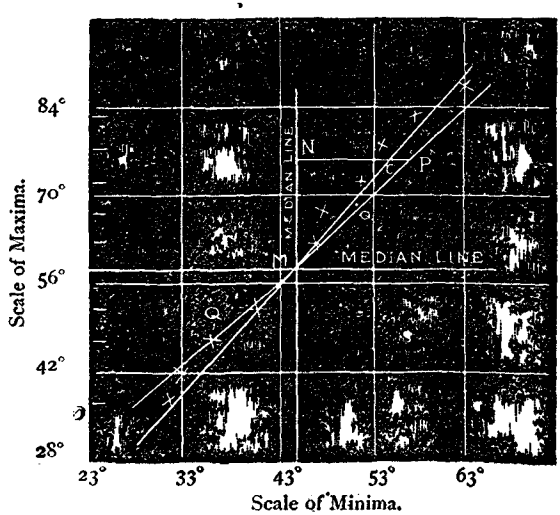
	Median	Quartiles		Distance between Quartiles
MINIMA	44½°	37°	51°	14°
MAXIMA	57½°	49°	67°	18°

To face page 323.

where the correlation is likely to be great, namely, the daily records of maxima and minima temperature recorded for 1898.*

We first make a rapid survey of the series, and notice that the minima range from about 23° to 63°, and the maxima from about 35° to 95°. Divide each of these ranges into, say, 10 equal parts, and draw up the framework of a table like the annexed. Turn through the records and enter the maximum and minimum for each day by a dot in the appropriate place; thus on 4th October the maximum was 61.3° and the minimum 51.3°; a dot should be put in the row 60°-64.9° under the heading 51°-54.9°. When all the dots are entered, replace them by their number in each square. The table shows the result for 358 days. If there is correlation, it will be found that the medians, or arithmetic averages, of each row form an orderly progression, and similarly for each column. These medians are roughly estimated and given in the table.

To test the correlation of the minima relative to the maxima a diagram is drawn. Choose scales so that the distance between the quartiles of the maxima (18°) shall be represented by the same length vertically, as represents the distance between the quartiles of the minima (14°) horizontally. Place crosses horizontally level with the middle points of the successive limits



* *Whitaker's Almanack*, 1899.

of maxima and vertically above the positions on the scale of the medians of the corresponding minima.

Now draw two lines. The first through the positions of the quartiles and median (Q_1, Q_2, M); this is the line of perfect correlation, and with the scales we have chosen is at 45° to the horizontal; draw another line through M,* passing as near as possible to all the crosses. Draw any horizontal line PCN intersecting the former lines as in the figure. The ratio of CN to PN is the coefficient of correlation. If the line CM passes through all the crosses and coincides with PM, the correlation is perfect. If CM is perpendicular to PM, there is perfect negative correlation. If CM is vertical there is no correlation. In the figure the ratio CN to PN is $\frac{4}{5}$. A rough test of the presence of correlation is to be obtained by noticing whether all the crosses above the median are on one side of NM and below the median on the other side.

There is a simple connection between the coefficient thus determined and that obtained by the previous formula. On the diagram the scales are so chosen that we replace $\frac{x}{\sigma_1}, \frac{y}{\sigma_2}$ by quantities ξ, η measured by equal units. Then if $(\xi_1\ \eta_1)\ (\xi_2\ \eta_2) \ldots$ are the positions of the 358 original pairs the line $y = rx$ can be shown to be that whose mean† distance from these points is a minimum when $r = \frac{\Sigma \xi\eta}{n}$, its value previously given. It is easily seen that in the figure the ratio $\frac{CN}{PN}$ is r_1, if $y = r_1 . x$ is the equation of a line through M referred to horizontal and vertical axes. Hence the line CM might be drawn from the original formula by taking $\frac{CN}{PN} = \frac{\Sigma \xi\eta}{n}$; in other words, we have here a graphic method of finding the coefficient of correlation.

Relation between the methods.

Calculating r roughly from the data $\sigma_1 = 12.7$, $\sigma_2 = 9.1$, $n = 358$, $\Sigma xy = 32130$, $r = \frac{32130}{358 \times 12.7 \times 9.1} = .8$ approx.; that is, we obtain approximately the same value by either method.

Mr Galton applied this method to the question of inheritance of stature. He found that the coefficient of correlation between the statures of children and of their (mid-) parents was $\frac{2}{3}$. That is if a group of (mid-) parents had an average stature x inches above (or below) the general average, the average for their sons was only $\frac{2}{3}\ x$ inches above (or below)

Regression.

* Perhaps the best way to do this is to rotate a line about M till the number of crosses on its two sides are equal.

† For the sense in which this word is to be taken, see Yule, *loc. cit.*, pp. 817-8.

the general average. This return towards the average is called in biological language "regression"; such equations as $y = r.\frac{\sigma_2}{\sigma_1}x$, $x = r.\frac{\sigma_1}{\sigma_2}y$ are the "equations of regression"; and the quantities $r.\frac{\sigma_2}{\sigma_1}$ and $r.\frac{\sigma_1}{\sigma_2}$ are the "coefficients of regression," the one when the sons of a group of parents of given height are compared with their parents, the other when the parents of a group of parents of given height are compared with their sons, where σ_1, σ_2 are the standard deviations in height for the parents and sons respectively.

Evolution.

There is an intimate relation between the law of error and biological theory. The law of error and other cognate laws give algebraic expression to the universal tendency to variation, whether we are dealing with any part of the social organism to whose measurement we have in this book limited statistics, or with any measurable organ of an animal or vegetable. The law of heredity can be only tested numerically by the theory of correlation; the effect of natural selection is easily considered with the help of the coefficient of regression. For if there is no selection, the distance from the general average of the mean stature of successive generations, descended from a group whose mean deviation was x, will be rx, r^2x . . . r^nx if r remains unchanged, a series whose terms rapidly tend to zero. If on the other hand a selection is made in each generation of those above the average, the divergence can be preserved and intensified. The discussion of this point would lead us too far afield.

Conclusion.

In this Second Part we have only discussed the elements of the subject, the theorems and formulæ which writers on statistics now assume. We have examined only the normal curve of error, postponing the asymmetrical curve of error to the appendix, and have not touched algebraic formulæ arising from different hypotheses, or correlation between more than two variables. In the region to which we have confined ourselves, however, we have had to deal with arguments of the same nature as are to be met with in the higher paths of statistics. The great difficulty which the student of economics encounters when dealing with the theory of error is the apparent slightness of relation between this theory and the facts with which he deals. This slightness is only apparent; it is because the theory has not, in the form he meets it, been carried far enough

to fit it to the very complex facts of human affairs that we do not get that exact correspondence we might desire. The theoretical distribution of error may be expected to underlie all phenomena, just as the attraction of gravity underlies the action of all machinery. We cannot explain the motion of machinery by gravity alone, we need to consider also other natural forces, not so easily measured as gravity; but still less can we explain that action if we ignore the force of gravity. It is hoped that the short treatment here given of the elements of so important a subject may make smoother the approach to a field of investigation where there is great promise of harvest but where the reapers are as yet few.

Note.—While this book has been in the Press, an article by Prof. Pearson has appeared in the *Philosophical Magazine*, July 1900, violently criticising the method adopted by most of his predecessors, who have investigated the applicability of the Law of Error to Statistics, that is to say, the method of first deducing the equation from *à priori* considerations, and then comparing the results with experiments. By means of a criterion of "fitting," which should be carefully studied, he shows that the chances that the statistics, with which Airy and Merriman illustrate the theory, would have arisen from random sampling are only .01423 and .000,00155 respectively on their hypothesis, and deduces "that the normal curve possesses no special fitness for describing errors or deviations such as arise either in observing practice or in nature." It is to be remarked on this, first that the investigation of two examples does not prove his case, secondly that his criticism does not apply to such curves as the asymmetrical curve of error treated exhaustively by Professor Edgeworth, and thirdly that the claim of the authors, whom he treats with such contempt, is not that the fit is exact, but "that the formula represents with all *practicable accuracy* the observed frequency" (Airy, quoted by Pearson) or "that the agreement is *very satisfactory*" (Merriman): thus the authors in question make no claim that the normal law is the complete explanation of the observed errors, but are satisfied with the approximation they found: it was not to be expected that the pioneers in the field should attain finality. By a similar process the law of gravitation might be treated with derision by criticising the experiments of an Attwood's machine, when the resistance of the air was not considered. Prof. Pearson has four constants in the curve by which he attains a close fit in his Illustration IV., and by increasing the number of his constants might obtain an absolute fit. With those developments of the normal curve of error, which depend on hypotheses very similar to those used by the earlier writers (see p. 303, *supra*, and Professor Edgeworth's recent contributions to the *Statistical Journal*), more constants are present, and there is every likelihood that equally close agreement may be found. The present author does not, however, wish to enter here into the controversy as to which is the best formula for classifying phenomena. His intention has been to follow in the beaten track, and there can be little doubt that the ordinary reader will prefer to find some

à priori justification for the unfamiliar theory that natural phenomena can be represented by the formulæ of algebraic probability, *pace* the author of *The Grammar of Science*, though he may recognise that the ultimate justification for the theory must be experience. There is no suggestion in this book that the whole of nature can be measured by the foot-rule of the normal curve of error; but yet that it may be a useful instrument has been shown by few people more conclusively than by Prof. Pearson himself.

In the following list will be found those books and articles relating to the subject of Part II. of this book which are most accessible and likely to be most useful to the English student. Further references to foreign authors and to earlier writers will of course be found in the works here mentioned :—

TODHUNTER, I.—*History of the Theory of Probability.* Especially Arts. 993-1002.

ENCYCLOPÆDIA BRITANNICA.—Article on *Probability.*

DICTIONARY OF POLITICAL ECONOMY (Palgrave's).—Article on *Law of Error.*

GALTON, F.—*Inquiries into Human Faculty and its Development.*

—— *Natural Inheritance.*

—— *Family Likeness in Stature.* Proc. of Royal Soc., 1886, 1888.

MERRIMAN, M.—*Method of Least Squares.*

CHAUVENET.—*Practical and Spherical Astronomy*, vol. ii., App.

EDGEWORTH, Prof. F. Y.—In the London, Edinburgh and Dublin *Philosophical Magazine and Journal of Science* (formerly issued under other similar titles, and known as the Literary and Philosophical Magazine), 5th series.

Vols. 21, 22, 23, 24, 25, 30. Various investigations and examples.

Vols. 34, 35, 36. Correlation.

Vol. 41. Asymmetrical law of error.

—— In the *Journal of the Royal Statistical Society*, 1886 and Jubilee Volume Methods of Statistics.

1888 and 1890. Chance in competitive examinations.

1893 and 1894. Correlation.

1895. Recent contributions (Pearson's) to theory.

1896, 1897, and 1898. Miscellaneous applications of the Calculus of Probabilities.

1899 and 1900. Representation of Statistics by Mathematical Formulæ.

—— Report of Committee of British Association on Monetary Standard, 1888.

—— Camb. Phil. Soc. Trans., 1885 and 1886. Merits of various means.

[The exact titles of the above articles may be found from the indexes of the volumes mentioned.]

PEARSON, Prof. K.—*The Chance of Death and other Essays.*

—— *The Grammar of Science*, chaps. x., xi.

—— *Contributions to Mathematical Theory of Evolution* in Transactions of Royal Society, 1894, 1895, 1896, and Stat. Soc. Journal, 1896, 1897.

—— *Probable errors of frequency constants.* Royal Soc. Trans., 1898.

—— *Criterion . . . of Deviations . . . in a Correlated System . . .* Phil. Mag., July 1900.

VENN, Dr J.—*The Logic of Chance.*

—— *Nature and Use of Averages.* Stat. Journal, 1891.

—— *Cambridge Anthropometry.* Journal of Anthropological Institute, Nov. 1888.

YULE, U.—*History of Pauperism.* Stat. Journal, 1896.
—— *Theory of Correlation.* Do. 1897.
—— *Changes in Pauperism.* Do. 1899.
—— *Association of Attributes in Statistics.* Royal Soc. Trans., 1900.
BOWLEY, A. L.—*Accuracy of an Average.* Stat. Journal, 1897.
SHEPPARD, W. F.—*On the calculation of the Average Square.* Stat. Journal, 1897.
—— *Use of Auxiliary Curves.* Stat. Journal, 1900.
—— *Normal Correlation.* Camb. Phil. Society, vol. xix.
—— *Normal Distribution and Correlation.* Royal Soc. Trans., 1898.

APPENDIX.

On the Measurement of an Unsymmetrical Group.

In the chapter on Averages, and in Section II. of Part II., we have developed a system of measurement which applies in its entirety only to a symmetrical group. It is true that the statement of the two quartiles and the median, or of any group of percentiles more than three in number, allows us to describe a group whether symmetrical or not; but the use of the "dispersion," p. 136, or of one of the quantities r, c, η, or ϵ of Part II., gives us a measurement of the relation of a group to its average, which is incomplete till we either know that the group is symmetrical or have a measure of its skewness.

The relative position of the arithmetic average and the median, or of the arithmetic average and the mode, or of the two quartiles and the median, clearly have some relation to the skewness.

The curve of error throws considerable light on these relations, and affords us several means of calculating a measure of skewness generally applicable to the description of groups.

If in the equation on p. 276, we put $x = z \times \sqrt{2pqn}$, thus making z finite, and then include terms of the order $n^{-\frac{1}{2}}$, we obtain that the chance of drawing $pn+x$ white balls

$$= Pe^{-z^2}\left[1 - 2j\left(z - \tfrac{2}{3}z^3\right)\right], \text{ where } j = \frac{q-p}{2\sqrt{2pqn}}.$$

Replacing z by $\frac{x}{c}$, where c^2 is (as before) $2pqn$, we have for the equation of the curve—

$$y = \frac{N}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}\left[1 - 2j\left(\frac{x}{c} - \frac{2}{3}\frac{x^3}{c^3}\right)\right],$$

where the unit of ordinates is so taken that the area of the whole curve is N, the number of observations.*

* An equivalent formula is given by Westergaard (*Die Grundzüge der Theorie der Statistik*, p. 70) from Laplace; see Todhunter's *History of the Theory of Probability*, Art. 993.

j is thus equal to $\frac{q-p}{2c} = \frac{q-\frac{1}{2}}{c} = \frac{\frac{1}{2}-p}{c}$.

The abscissa of the centre of gravity is still zero, and the error of mean square $\frac{c}{\sqrt{2}}$.

j has been so chosen that it is equal to the mean cube of error, divided by c^3 to make it independent of the unit.*

Writing the above equation in its integral form, and putting Y_τ for the area standing on the line CX, where C is the origin, and X a point on the axis with abscissa x, and $\tau = \frac{x}{c}$, we have

$$Y_{\pm\tau} = \frac{N}{\sqrt{\pi}}\left\{\int_0^\tau e^{-\tau^2}.d\tau \mp \frac{j}{3}\left(1 - e^{-\tau^2}\frac{6}{1-2\tau^2}\right)\right\}$$
$$= N\{F(\tau) \mp j.f(\tau)\}.$$

The double sign is due to the fact that jc must be taken as $\overline{p-\frac{1}{2}}$ on the p side of the centre.

F (τ) is tabulated on p. 281, $f(\tau)$ on p. 332.

A tabulation of Y_τ for certain values of j will be found in the *Statistical Journal* for July 1902.†

j should be calculated by a method similar to that used for calculating c on p. 290; the cubes of errors being taken instead of the square, and the result divided by c^3. An example is given below.

Professor Edgeworth has shown that, with j so calculated, the equation above given is a correct expression for the curve of error resulting from the hypotheses stated on p. 303. The normal curve of error, discussed in sections I. to V., is the first approximation, when j is taken as zero. If j is not small, this second approximation breaks down not far from the centre; but if j is not greater than .3, a great part of the curve is correctly given.

The curve, whose equation contains j, intersects the curve obtained by putting j = 0, when $x = 0$ or $\pm c.\sqrt{\frac{3}{2}}$. This is seen on the accompanying diagram.

---

* The mean cube of error $= \dfrac{\int_{-\infty}^{+\infty} yx^3.dx}{\int_{-\infty}^{+\infty} y.dx}$ which can be verified as equal to j.

† There $f(\tau)$ does not include $\frac{1}{3\sqrt{\pi}}$, and its sign is reversed.

The median, O, can be shown to be approximately at the point $\tau = -\frac{1}{3}j$; the mode, M, at the point $\tau = -j$; the quartiles, Q_1, Q_2, at $\tau = \pm\rho + \frac{j}{3}(2\rho^2 - 1) = \pm .477 - .182j$, where ρ has the same meaning as on p. 292, and is written for .4769.

Thus j may be calculated from observing any two of the above positions (when c is known, and the observations reduced to unit c).

If we do not know c, the positions of these points* are given by

$$CO = -\tfrac{1}{3}jc$$
$$CM = -jc \dagger$$
$$OQ_2 = \rho c + \tfrac{2}{3}\rho^2 jc$$
$$OQ_1 = \rho c - \tfrac{2}{3}\rho^2 jc,$$

where C is the position on the axis of the centre of gravity.

Thus the observation of the median and two quartiles gives us a rapid means of calculating j and c; for $c = \frac{OQ_1 + OQ_2}{.954}$ and $j = \frac{OQ_2 - OQ_1}{OQ_2 + OQ_1} \times 3.14$.

This method of calculating j bears nearly the same relation to the longer method given below, as the use of the average error or probable error (see p. 292) does to the use of the error of mean square.

j, calculated by any of these methods, is the measure of the skewness of the curve, which it was our object in this section to find.

Even when a group has no very close connection with the first or second approximation to the curve of error, it seems probable that c and j, calculated by the methods which for the particular group seem the least liable to chance disturbance, are the best single measures of the grouping about the average and the skewness, that we can devise.

The following example shows the method of calculating c and j by means of mean square and mean cube of error.‡

---

* These equations are due to Professor Edgeworth; their use is investigated in a joint article by him and the present author in the *Statistical Journal*, July 1902.

† Hence $CM = \frac{1}{2}(p - q)$.

‡ Other examples are to be found in the *Statistical Journal*, *loc. cit.*

VALUES OF $f(x)$ FOR DIFFERENT VALUES OF x, WHERE

$$f(x) = \frac{1}{3\sqrt{\pi}}[1 - e^{-x^2}(1 - 2x^2)].$$

x.	$f(x)$.	x.	$f(x)$.	x.	$f(x)$.	x.	$f(x)$.	x.	$f(x)$.
.00	.000	.36	.066	.71	.189	1.06	.264	1.41	.264
.01	.000	.37	.069	.72	.192	1.07	.265	1.42	.264
.02	.000	.38	.072	.73	.195	1.08	.266	1.43	.263
.03	.001	.39	.076	.74	.198	1.09	.267	1.44	.262
.04	.001	.40	.079	.75	.201	1.10	.268	1.45	.262
.05	.001	.41	.082	.76	.204	1.11	.268	1.46	.261
.06	.002	.42	.086	.77	.207	1.12	.269	1.47	.260
.07	.003	.43	.090	.78	.210	1.13	.270	1.48	.259
.08	.004	.44	.093	.79	.213	1.14	.270	1.49	.258
.09	.005	.45	.097	.80	.216	1.15	.270	1.50	.257
.10	.006	.46	.100	.81	.218	1.16	.271	1.52	.256
.11	.007	.47	.104	.82	.221	1.17	.271	1.54	.254
.12	.008	.48	.107	.83	.224	1.18	.271	1.56	.252
.13	.009	.49	.111	.84	.226	1.19	.272	1.58	.250
.14	.011	.50	.115	.85	.229	1.20	.272	1.60	.248
.15	.013	.51	.118	.86	.231	1.21	.272	1.62	.246
.16	.014	.52	.122	.87	.233	1.22	.272	1.64	.244
.17	.016	.53	.126	.88	.236	1.23	.272	1.66	.242
.18	.018	.54	.129	.89	.238	1.24	.272	1.68	.240
.19	.020	.55	.133	.90	.240	1.25	.272	1.70	.238
.20	.022	.56	.137	.91	.242	1.26	.272	1.72	.236
.21	.024	.57	.140	.92	.244	1.27	.271	1.74	.234
.22	.026	.58	.144	.93	.246	1.28	.271	1.76	.232
.23	.029	.59	.148	.94	.248	1.29	.271	1.78	.230
.24	.031	.60	.151	.95	.249	1.30	.271	1.80	.228
.25	.033	.61	.155	.96	.251	1.31	.270	1.82	.227
.26	.036	.62	.158	.97	.253	1.32	.270	1.84	.225
.27	.039	.63	.162	.98	.254	1.33	.269	1.86	.223
.28	.041	.64	.165	.99	.256	1.34	.269	1.88	.221
.29	.044	.65	.169	1.00	.257	1.35	.268	1.90	.220
.30	.047	.66	.172	1.01	.259	1.36	.268	1.92	.218
.31	.050	.67	.176	1.02	.260	1.37	.267	1.94	.217
.32	.053	.68	.179	1.03	.261	1.38	.267	1.96	.215
.33	.056	.69	.182	1.04	.262	1.39	.266	1.98	.214
.34	.059	.70	.186	1.05	.263	1.40	.265	2.00	.212
.35	.062							2.05	.209

x.	$f(x)$.	x.	$f(x)$.	x.	$f(x)$.
2.10	.206	2.50	.192	2.90	.189
2.20	.201	2.60	.191	Above	
2.30	.197	2.70	.190	2.90	.188
2.40	.194	2.80	.189		

The data are the number of school children of various ages in the sixth grade in the public schools of St Louis, U.S.A. On the axis of abscissæ (z), are measured the ages. One year is taken as the unit, and $10\frac{1}{2}$ years as the origin from which the ages are measured. The number in each age group is denoted by y. μ_1', μ_2', μ_3' are the average first, second, and third powers of the deviations from the origin. μ_1, μ_2, μ_3 corresponding quantities measured from $\bar{z}$ ($=\mu_1'$) the centre of gravity.

One student over 19 years is excluded.

Ages. Years.	Numbers at each Age. y	Corresponding Values of. z	yz	yz^2	yz^3.
10–11	26	0	0	0	0
11–12	201	1	201	201	201
12–13	673	2	1346	2692	5384
13–14	1001	3	3003	9009	27027
14–15	739	4	2956	11824	47296
15–16	310	5	1550	7750	38750
16–17	80	6	480	2880	17280
17–18	13	7	91	637	4459
18–19	1	8	8	64	512
	$N=3044=\Sigma y$		9,635	35,057	140,909

$$\bar{z}=\mu_1'=\frac{\Sigma yz}{N}=\frac{9635}{3044}=3.165.$$ Average age is $10\frac{1}{2}+\bar{z}=13.665$ (years).

$$\mu_2'=\frac{\Sigma yz^2}{N}=11.517 \qquad \mu_3'=\frac{\Sigma yz^3}{N}=46.291.$$

$$\mu_2=\frac{\Sigma y(z-\bar{z})^2}{N}=\frac{\Sigma yz^2}{N}-2\bar{z}\frac{\Sigma yz}{N}+\bar{z}^2\frac{\Sigma y}{N}=\mu_2'-\bar{z}^2=1.500$$

$\sqrt{\mu_2}$ is then the error of mean square; but Mr Sheppard * has shown that when, as in this example, the numbers are supposed to be at the centres of their groups (at $10\frac{1}{2}$, $11\frac{1}{2}$ years, &c.), instead of being scattered through them, it is proper to subtract $\frac{1}{12}$ from μ_2, as calculated above.

Then $c=\sqrt{2(\mu_2-\frac{1}{12})}=1.683$ (years).

$$\mu_3=\frac{\Sigma y(z-\bar{z})^3}{N}=\mu_3'-3\mu_2'\bar{z}+2\bar{z}^3=.347$$

Then $j=\frac{\mu_3}{c^3}=.0728$.

* See *Statistical Journal*, Sept. 1897.

Putting τ for $\frac{x}{c}$, we have the following scheme for finding the values of Y_τ between the average age and the years 11, 12 . . .

Ages.	x.	τ.	$F(\tau)$.	$f(\tau)$.	$jf(\tau)$.	$F(\tau)$ * $\pm jf(\tau)$	Differences of Col. 7.	Numbers: Calculated Col. 8 Mult. by N.	Numbers: Actual.	Differences or Errors.
—	— ———	∞	.500	.188	.014	.514	.009	27	26	1
11 years	−2.665	−1.583	.487	.249	.018	.505	.067	204	201	3
12 ,,	−1.665	− .989	.419	.256	.019	.438	.220	670	673	3
13 ,,	− .665	− .395	.212	.077	.006	.218	.327	995	1001	6
14 ,,	+ .335	+ .199	.111	.022	.002	.109	.245	746	739	7
15 ,,	+1.335	+ .793	.369	.214	.015	.354	.101	307	310	3
16 ,,	+2.335	+1.387	.475	.266	.020	.455	.026	79	80	1
17 ,,	+3.335	+1.981	.497	.214	.016	.481	.005	15	13	2
18 ,,	+4.335	+2.575	.500	.191	.014	.486	0	0	1	1
19 ,,	+5.335	+3.170	.500	.188	.014	.486				
—	—	∞	.500	.188	.014	.486				27

The sum of the difference is thus 27 on a total of 3044, or about .9 per cent.

The misfit without the j term is more than five times as great, viz., 5.1 per cent.

The median is at $(13.665 - \frac{1}{3} j c)$ years, that is at 13.624 years.

The mode is at 13.543 years.

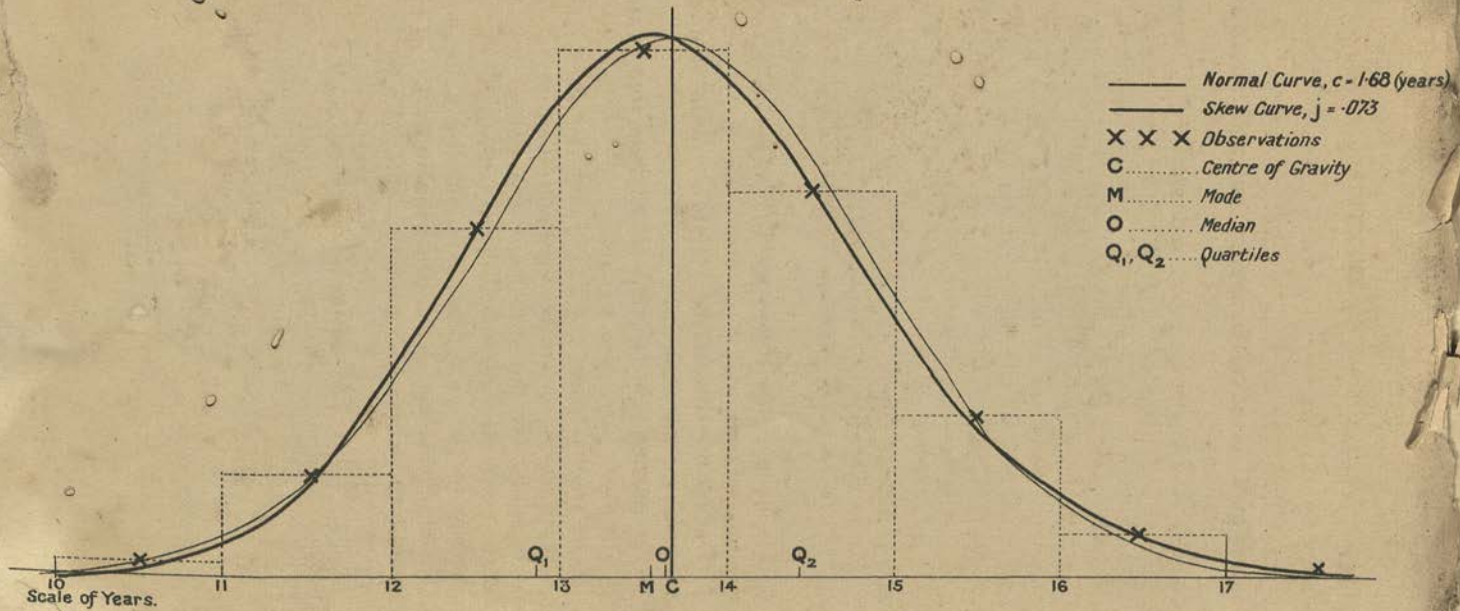
The normal curve ($c = 1.683$) and the skew curve ($c = 1.683$, $j = .0728$) are shown on the adjoining diagram, together with the observations. If the agreement was perfect the areas of the rectangles formed by the ordinates through the points corresponding to the years, the axis, and the parallels through the calculated positions, would equal the areas cut off from the skew curve by the same ordinates.

* Additions when τ is negative.

To face page 334

EXAMPLE OF THE SKEW CURVE OF ERROR.

GROUPING OF SCHOOL CHILDREN IN VITH GRADE OF S. LOUIS PUBLIC SCHOOLS ACCORDING TO AGE.



INDEX.

(References to definitions are printed thus:—109.)

ACCURACY, 199-214.
Age, 29, 147, 251, 312, 333.
Agricultural Wages, 50-52, 97-103, 109, 110, 115-117.
Arithmetic Average or Mean, 107-110, *109*, 125, 126, 128, 129, 130, 136, 221, 329; error in, 204, 306.
Asymmetry, see *Skewness*.
Average error, *285*, *292*, 329, 331.
Average wage, 6, 11.
Average: Precision of, 305-308.
Averages; 7, 89, 89, 92, 95, 107-130, *130*, 133-140, 143, 214, 264; see *Arithmetic Average*, *Median*, *Mode*, *Weighted Average*.

BERTILLON, Dr J., 17, 129, 130, 158.
Bias, 118.
Biassed errors, 209-214, 219.
Bibliography: of Interpolation, 258; of Law of Error, 327.
Binomial Expansion or Theorem, 265, 272-277, 288, 291, 301-2, 306.
Births, 287.
Blank Forms, 18, 19, 26, 27, 30, 37, 42, 46, 63; specimens of, 23, 35, 36, 45, 48, 51, 52, 54-58, 65, 67, 69.
Boole, 242, 247, 251.
Booth, C., 9, 27, 32, 78-80, 123, 158, 251.
Bortkewitsch, Dr, 302.

CARTOGRAMS, 156-158.
Census: Population, 10, 11, 23-32, 63, 78-81, 82, 99, 233.
Census: Wage, 11, 12, 33-40, 63, 87, 92-96, 114, 125, 233.
Chance, 266, 267; see *Probability*.
Changes in Wages, 54-58, 61, 97-103.
Chauvenet, 303.
Coefficient: of Correlation, *318*; of Probability, *299*; of Regression, *325*.
Coefficients: Statistical, *129*, 130, 296, 299.
Collection of material, 17, 18.
"Combinational" Groups, 299.
Comparison: Accuracy of, 206, 212, 305.
Comparisons of Series, 168-177, 192-194; see *Correlation*.
Consumption: Index No. of, 228.
Correlation, *316*, 317-325.
Cotton: wages, 39, 95, 96, 114; trade, 164-167.
Curve of Error, see *Error, Law of*.
Curves of Frequency, *303*.
Cycles of Trade, 153, 181.

DARWIN, G. H., 256.
De Morgan, 242, 247.
Deciles, *124*, 125-128, 133, 136, 144.
Demography, 6, 7, 23.
Deviations, 126.
Deviation: Standard, 285, 292; Quartile, 282.
Diagrams, 19, 88, 143-196.
Dispersion, *136*, 140, 329.

EARNINGS, 37.
Economist, The, 11, 214, 221, 223-224.
Edgeworth, Prof. F. Y., 118, 187, 253, 254, 257, 262, 285, 299, 303, 307, 313, 315.
Employment, see *Unemployment*.
Equation of Regression, *325*.
Equation of curve of error: normal, 283; skew, 329.
Error, *201*, 203-214, 330.
Error, Curve of, 261, 269-292, 309-310, 311, 329-334; *Law of*, 5, 267, 303-315.
Error of mean cube, 330, 331.
Error of mean square, *285*, 286, 290, *292*, 330, 331, 333; see *Standard Deviation*.
Error, probable, see *Probable Error*.
Exports, see *Foreign Trade*.

FACILITY CURVES, *303*.
Fluctuation, *285*, *291*, 298.
Foreign Trade, 11, 63-70, 148, 151-4, 170-1, 174-7, 188-191, 221-3, 250, 320-2.
Forms of Inquiry, see *Blank Forms*.
Fox, W., 100, 124.
French Wages, 39, 40.
Frequency Curves, *303*.

GABAGLIO, Prof. A., 156.
Galton, F., 89, 126-8, 322, 324.
Geometric Mean, *128*, 129, 221-3.

Giffen, Sir R., 10, 70, 151-4.
Graphic Method, see *Diagrams;* of Interpolation, 238-9.
Great Numbers, 8, 263-4.

HISTORICAL DIAGRAMS, 159-167.
Hours of work, 54, 57, 58.

IMPORTS, see *Foreign Trade.*
Index-Numbers, 111-2, 190, 217-29.
Interpolation, 19, 233-58.

JEVONS, W. S., 128, 178.

LABOUR COMMISSION, 37.
Labour Department, 10, 41-62, 63, 97-103.
Labour Gazette, The, 12, 41, 44, 46, 48, 50, 58, 60, 239.
Labour Statistics, 10.
Large Numbers, 4.
Least Squares, 5, 177.
Le Play, P. G., 7.
Levasseur, P. E., 156.
Levi, Leone, 9.
Lexis, Prof. W., 263, 280, 297, 298, 299.
Logarithmic Curves, 188-196.
Logarithms, Table of, 195-6.
Luck, 293.

MAKEHAM'S Formula, 257.
Marriage Rate, 108, 174-7, 193-4, 320-2.
Maximum Ordinate, *119;* see *Mode.*
Mean of Errors; see *Average Error.*
Median, 95, 117, 123, *124,* 125-8, 130, 133, 136, 138, 144, 154-6, 221, 224, 290, 323, 329, 331, 334; determination of, 127, 155, 252.
Merriman, M., 256, 284, 303.
Method: of Least Squares, 5, 177; Statistical, 4, 7, 17-20.
Mode, *119,* 118-124, 130, 133, 136, 144, 155-6, 329, 331, 334; determination of, 155, 252.
Modulus, *283-5,* 288, 289, 290, *291,* 299, 303-9, 312, 313, 314, 329-334; probable errors of, 319.

NORMAL CURVE OF ERROR, 277, 303, 304, 309, 330.

OCCUPATION, 29, 82, 99, 101.
Official Statistics, 9, 10, 213.

PEARSON, Prof. Karl, 5, 316, 318, 319, 326.
Percentiles, 124, 329.
Periodic Figures, 178-87, 240, 315.
Population, 28; see *Census.*
Poynting, Prof. J. H., 181.
Precision, 201, *285,* 286, *292,* 309; scale of, 273; of an average, 305-8; see *Modulus.*
Prices, see *Index-Numbers.*
Probable Error, *281,* 290, *292,* 305, 306, 307, 308, 320, 329, 331; of coefficient of regression, 319; of modulus, 319.
Probability, 20, *269.*
Purchasing Power, see *Index-Numbers.*

QUARTILES, 95, *124,* 125-8, 133, 134, 136, 144, 155, 290, 313, 320, 329, 331.
Quartile Deviation, 282.
Questions, 24.
Quetelet, A., 118-9, 124, 272-3, 278, 280, 284, 285.

REGRESSION, 324, 325.
Retail Prices, 11; Index-Number of, 225-9.
Revenue, 159-61.

SAMPLES, 20, 219, 224, 225, 308-313.
Sauerbeck, A., 190, 223-4.
Sheppard, W. F., 255.
Skew curve, 329, 334.
Skewness, 277, 329, 331.
Small numbers, 301.
Smoothing, 151-6.
Standard Deviation, *285, 298.*
Statistics, *3, 4, 7,* 262; official, 9, 10.
Statistical Abstract, The, 10.
Statistical Coefficients, *129,* 130, 296, 299.
Stirling's formula, 280.
Strikes, 51, 54-62.
Summary, 17, 19.
"Symptomatic" Series, 240, 315.

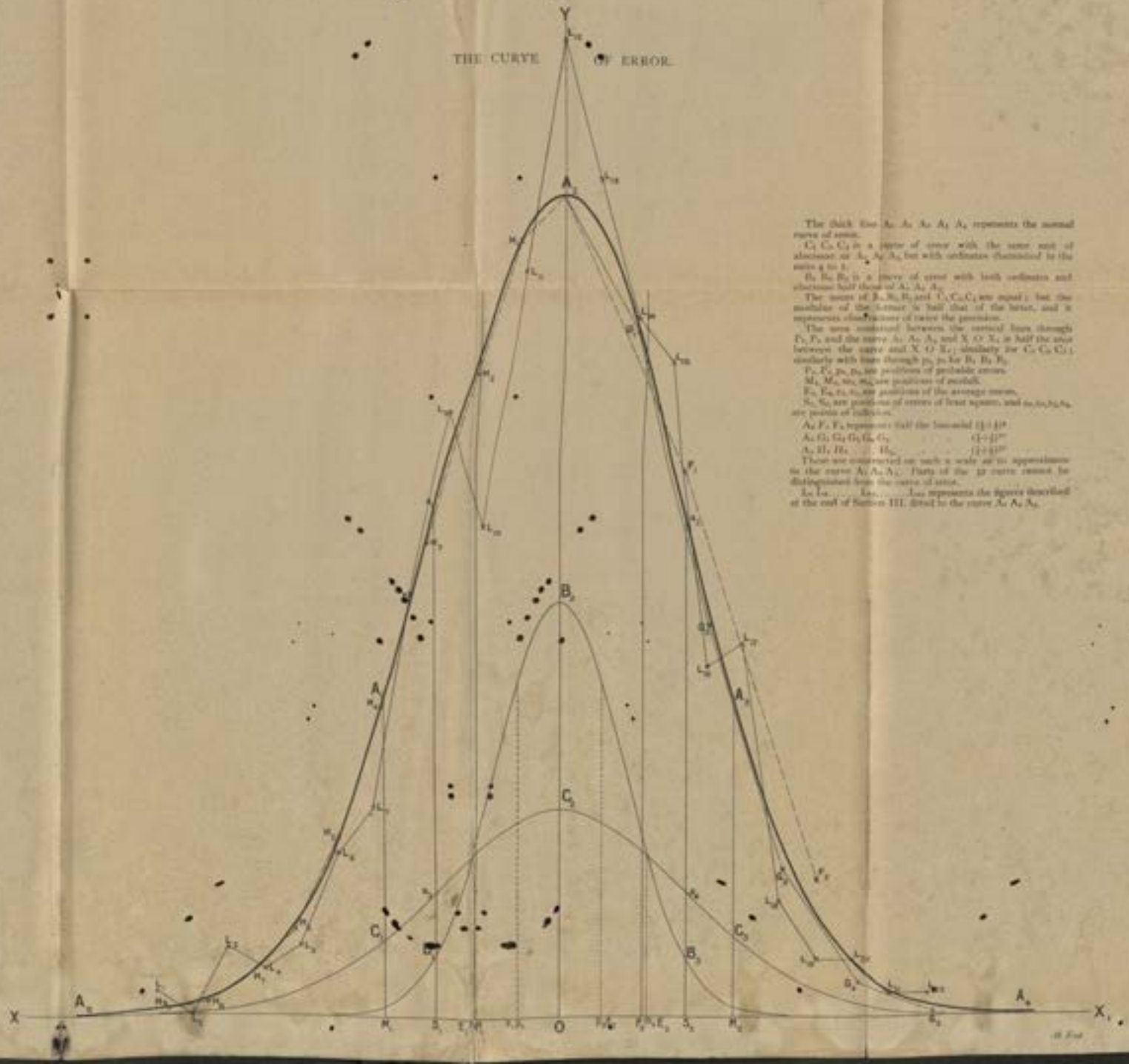
TABULATION, 17, 18, 24, 73-103, 133-140.
Tellers, 3, 25, 26, 28, 31.
Trade Unions, 42, 46, 53, 81.
Type, 6, 124, 130.

UNBIASSED ERRORS, 209-14, 219.
Unemployed, 40, 41, 42, 45, 46, 52, 178-187, 192-4.

VENN, Dr J., 262, 266.

WAGE CENSUS, see *Census.*
Wage Statistics, 11, 87-92, 120, 134-6, 144-6, 250, 310.
Wages, 54, 57, 58, 61, 149, 150; see *Agriculture.*
Weighted Average, *111,* 112-8; errors in, 205, 207, 214, 219-22, 304.
Westergaard, Prof. H., 287, 329.
Wheat, 161-3, 174-7, 186, 320-2.
Wholesale Prices, 11; see *Index-Numbers.*
Wood, G. H., 193, 228.

YULE, U., 282, 317, 319.



The thick line $A_0 A_1 A_2 A_3 A_4$ represents the normal curve of error.

$C_1 C_2 C_3$ is a curve of error with the same unit of abscissæ as $A_0 A_1 A_2$ but with ordinates diminished in the ratio 4 to 1.

$B_1 B_2 B_3$ is a curve of error with both ordinates and abscissæ half those of $A_1 A_2 A_3$.

The areas of $B_1 B_2 B_3$ and $C_1 C_2 C_3$ are equal; but the modulus of the former is half that of the latter, and it represents closer shooting of twice the precision.

The area contained between the vertical lines through P_1, P_2 and the curve $A_1 A_2 A_3$ and $X O X_1$ is half the area between the curve and $X O X_1$; similarly for $C_1 C_2 C_3$; similarly with lines through p_1, p_2 for $B_1 B_2 B_3$.

P_1, P_2, p_1, p_2, are positions of probable errors.

M_1, M_2, m_1, m_2, are positions of median.

E_1, E_2, e_1, e_2, are positions of the average errors.

S_1, S_2, are positions of centre of least squares, and s_1, s_2, s_3, s_4, are points of inflexion.

$A_0 F_1 F_2$ represents (nearly) the binomial $(\frac{1}{2}+\frac{1}{2})^{10}$.

$A_0 G_1 G_2 G_3 G_4 G_5$ ... $(\frac{1}{2}+\frac{1}{2})^{20}$.

$A_0 H_1 H_2$ ... H_{10} ... $(\frac{1}{2}+\frac{1}{2})^{40}$.

These are constructed on such a scale as to approximate to the curve $A_0 A_1 A_2$. Parts of the 20 curve cannot be distinguished from the curve of error.

$L_0 L_1$ ... L_{10} ... L_{20} represents the figure described at the end of Section III. fitted to the curve $A_0 A_1 A_2$.

00029224

Digitized with financial assistance from the
Government of Maharashtra
on 26 June, 2018